Ricardo Almeida Silva

Master Degree in Computer Science

Enhancing Exploratory Analysis across Multiple Levels of Detail of Spatiotemporal Events

Thesis submitted in partial fulfillment
of the requirements for the degree of

Doctor of Philosophy in
Computer Science

Adviser: João Moura Pires, Assistant Professor,
NOVA University of Lisbon

Examination Committee

Chairperson:	Professor Nuno Manuel Robalo Correia
Raporteurs:	Associate Professor David Taniar
	Full Professor Jerome Bernard Gensel
Members:	Associate Professor Maribel Yasmina Santos
	Assistant João Moura Pires

FACULDADE DE
CIÊNCIAS E TECNOLOGIA
UNIVERSIDADE NOVA DE LISBOA

June, 2017

Enhancing Exploratory Analysis across Multiple Levels of Detail of Spatiotemporal Events

This document was created using the (pdf)LaTeX processor, based in the "unlthesis" template[1], developed at the Dep. Informática of FCT-NOVA [2].
[1] https://github.com/joaomlourenco/unlthesis [2] http://www.di.fct.unl.pt

To you...

Acknowledgements

Os meus agradecimentos vão para todos aqueles que tornaram possível a realização desta dissertação, ou que de alguma forma tiveram impacto no resultado da mesma.

Gostaria de agradecer ao meu orientador, Professor João Moura Pires, em primeiro lugar, por ter apostado em mim. Em segundo, agradeço pela sua orientação e por todas as discussões interessantes e conselhos que me colocaram na direcção correcta. Em terceiro, agradeço todo o conhecimento que me foi transmitido. E por último, pelo seu apoio incondicional nos mais variados assuntos, muitas vezes não relacionados com aspectos com a dissertação mas que foram fundamentais para a conclusão desta.

Para a Professora Maribel Yasmina Santos, que desde o primeiro dia sempre esteve disponível. Os seus comentários pertinentes melhoraram, certamente, a qualidade desta dissertação. Para o Professor Nuno Datia, que não só teve uma participação activa no desenvolvimento do protótipo SUITE-VA, como também os seus comentários pertinentes moldaram a dissertação para o que ela é hoje.

Agradeço também ao Rui Leal que me ajudou num dos períodos mais dificeis. Esteve presente no desenvolvimento do primeiro protótipo, cujo trabalho teve contribuições importantes para a dissertação.

Um agradecimento para o Professor Bruno Martins, Professor Carlos Damásio, Professor Fernando Birra, e Professor Alferes. A todos, agradeço os comentários, críticas, sugestões e caminhos apontados.

Para todos os amigos que, de um modo ou outro, me ajudaram, em especial para aqueles que me acompanharam durante este percurso: Ana Sofia Gomes, Diogo Cardoso, Duarte Godinho, Daniela Gortan, Jorge Sousa, Lara Nascimento, Luísa Lourenço, Pedro Estrela, Sinan Elgimez, Jorge Costa, Miguel Domingues, Miguel Lourenço, Mário Pires, and many others.

Finalmente, um agradecimento aos meus pais por me sempre apoiarem incondicionalmente a terminar o doutoramento, mesmo quando parecia impossível.

Mas apesar do apoio de todas as pessoas mencionadas, o maior chegou da pessoa mais próxima de mim. Obrigado Carla, que tiveste um papel fundamental no equilibrio emocional, por toda a compreensão face à ausência, abdicação da vida pessoal e familiar, por me motivares nos momentos de frustração e celebrares os sucessos. A ti, um eterno Obrigado.

Abstract

Crimes, forest fires, accidents, infectious diseases, human interactions with mobile devices (e.g., tweets) are being logged as spatiotemporal events. For each event, its spatial location, time and related attributes are known with high levels of detail (LoDs). The LoD of analysis plays a crucial role in the user's perception of phenomena. From one LoD to another, some patterns can be easily perceived or different patterns may be detected, thus requiring modeling phenomena at different LoDs as there is no exclusive LoD to study them.

Granular computing emerged as a paradigm of knowledge representation and processing, where granules are basic ingredients of information. These can be arranged in a hierarchical alike structure, allowing the same phenomenon to be perceived at different LoDs. This PhD Thesis introduces a formal Theory of Granularities (ToG) in order to have granules defined over any domain and reason over them. This approach is more general than the related literature because these appear as particular cases of the proposed ToG. Based on this theory we propose a granular computing approach to model spatiotemporal phenomena at multiple LoDs, and called it a granularities-based model. This approach stands out from the related literature because it models a phenomenon through statements rather than just using granules to model abstract real-world entities. Furthermore, it formalizes the concept of LoD and follows an automated approach to generalize a phenomenon from one LoD to a coarser one.

Present-day practices work on a single LoD driven by the users despite the fact that the identification of the suitable LoDs is a key issue for them. This PhD Thesis presents a framework for **SU**mmariz**I**ng spatio**T**emporal **E**vents (SUITE) across multiple LoDs. The SUITE framework makes no assumptions about the phenomenon and the analytical task. A Visual Analytics approach implementing the SUITE framework is presented, which allow users to inspect a phenomenon across multiple LoDs, simultaneously, thus helping to understand in what LoDs the phenomenon perception is different or in what LoDs patterns emerge.

Keywords: Spatiotemporal events, granularity, level of detail, visual analytics

Resumo

Crimes, incêndios florestais, doenças infecciosas estão a ser registados como eventos espácio-temporais. Para cada evento, a sua localização espacial, tempo e atributos relacionados são conhecidos com grandes níveis de detalhe (NdDs). O NdD de análise tem um papel fundamental na percepção dos fenómenos. De um NdD para outro, alguns padrões podem ser facilmente perceptíveis ou diferentes padrões podem ser detectados, requerendo que os fenómenos sejam modelados a diferentes NdDs, uma vez que não existe apenas um NdD para os estudar.

A computação granular emergiu como o paradigma de representação de conhecimento e processamento, onde grânulos são ingredientes básicos de informação. Estes podem ser organizados numa estrutura hierárquica, permitindo que o mesmo fenómeno seja observado a diferentes NdDs. Esta dissertação introduz uma teoria formal de granularidades (TdG) de modo a ter grânulos definidos sobre qualquer domínio e a poder raciocinar sobre eles. Esta abordagem é mais geral do que a literatura relacionada porque as propostas da literatura mostraram-se casos particulares da TdG proposta. Com base nesta, uma abordagem de computação granular é proposta para modelar fenómenos espácio-temporais a múltiplos NdDs, designada de modelo baseado em granularidades. Esta abordagem destaca-se da literatura por modelar um fenómeno através de declarações em vez de apenas utilizar os grânulos para modelar entidades abstractas do mundo real. Além disso, formaliza o conceito de NdD e segue uma abordagem automática para generalizar o fenómeno de um NdD para outro menos detalhado.

As práticas actuais trabalham num NdD conduzido pelos utilizadores apesar da identificação dos NdDs apropriados ser um problema chave. É apresentada uma framework para sumarizar eventos espácio-temporais em múltiplos NdDs. Esta framework não faz qualquer assumpção sobre o fenómeno e a tarefa analítica. É apresentada uma abordagem de visualização analítica, que permite aos utilizadores inspeccionar um fenómeno em múltiplos NdDs, simultaneamente, ajudando a entender em quais NdDs a percepção do fenómeno se distingue ou em que NdDs emergem padrões.

Palavras-chave: Eventos espácio-temporais, granularidade, nível de detalhe

Contents

List of Figures

List of Tables

Chapter 1

Introduction

With the widespread adoption of location aware devices, organizations are gathering data (Committee et al. 2013), concerning information about geographic location and time (Li et al. 2016) at incredible rates. These data are usually called spatiotemporal data and contain information about natural phenomena or human activities occurring on or near the surface of the Earth like telecommunications, social networks, transport, health, meteorology and agriculture, among many others.

Many phenomena like crimes[1], storms[2], forest fires[3], infectious diseases (Gabriel et al. 2013), traffic accidents[4], social networks (e.g., Twitter) are being logged as a collection of spatiotemporal events at high levels of detail (LoDs).

A spatiotemporal event is a happening occurred in space and time (Yuan and Hornsby 2007). For example, *homicide((41.8780377, -87.6294422), 09/05/2015 20:00, 2)* stands for a homicide occurred on the latitude and longitude coordinates (41.8780377, -87.6294422) that happened at eight o'clock resulting in two victims; *fire((42.013990, -8.454387), 27/07/2016 14:30, 130)* describes a forest fire that has started at coordinates (42.013990, -8.454387) on 27th July 2016 at 14:30 hours leading to 130 hectares of burnt forest area. This way, spatiotemporal events could be described as data with the following structure: $event(S, T, A_1, \ldots, A_N)$ where S describes the geographic location of the event, T specifies the time moment, and $A_1, \ldots, A_N$ are attributes detailing what has happened.

[1] Crimes in City of Chicago: `http://data.cityofchicago.org/`
[2] Storm events in USA: `http://www.spc.noaa.gov/wcm/`
[3] Forest fires in Portugal: `http://www.icnf.pt/portal/florestas/dfci`
[4] Traffic accidents in USA: `ftp://ftp.nhtsa.dot.gov/fars/`

Datasets of spatiotemporal events embody the spatiotemporal dynamics of phenomena that includes data attributes changing over time or establishing several relationships or interactions with the surrounding environment. Underlying the complexity/dynamism inherent to spatiotemporal events, there might be hidden patterns to be uncovered (Miller and Han 2009).

Patterns are non-uniform distributions of events occurring in space or/and in time that reveal the underlying structure of a phenomenon (Mennis and Guo 2009). The appearance of crime hotspots in certain city areas is an example of a spatial pattern; hotspots of robberies near residential areas and hotspots of murders near town bars is an example of a spatial pattern with correlation between the attributes crime type and neighborhood type. An increase in the number of traffic accidents during the summer in every year is an example of a temporal pattern. The occurrences of tornadoes in particular spatial regions and in particular periods of the year, or the contagion of a disease are examples of spatiotemporal patterns.

Understanding patterns can be important for the decision-making of several organizations. For example, in public safety (Leipnik and Albert 2003), crime analysts are interested in discovering spatiotemporal hotspots of crime events in order to effectively allocate police resources. Epidemiologists (Ostfeld et al. 2005) need to understand spatiotemporal patterns from disease events so that the officials can allocate resources to limit its spreading. In what concerns the environment, state officials aim to understand spatiotemporal patterns of wildfire occurrences so that optimal firefighting resources and development projects can be placed in appropriate areas (Hering et al. 2009).

Visual Analytics (VA) aims at extracting patterns from data through smart combination of automatic algorithms and interactive visualization (Thomas and Cook 2006). By relying on human capabilities such as perception and domain knowledge, VA lets users to interactively explore the data and generate hypotheses while leveraging methods from knowledge discovery, data mining, artificial intelligence, statistics and mathematics.

Over the last years, several VA approaches have been developed that allow us to explore and analyze datasets of spatiotemporal events (Roth et al. 2010; MacEachren et al. 2011; Chae et al. 2012; Andrienko et al. 2013; Lins et al. 2013; Cho et al. 2016; Robinson et al. 2016). For example, the GeoVista research center developed CrimeViz (Roth et al. 2010) to study crimes and Senseplace (MacEachren et al. 2011) to support crisis management through the tweets posted; Lins et al. 2013 developed an approach to analyze numerous quantities of spatiotemporal events that was used to explore events about crimes, social networks, among others; Cho et al. 2016 developed an approach, called VAiRoma, for users to gain knowledge about places and events related to Roman history. An overview of the above mentioned tools' interfaces can be seen in Figure 1.1.

In general, the VA approaches developed, aiming at exploratory analysis of spatiotemporal events, use interactive visualizations (Van Ho et al. 2012) like maps/thematic maps, time series, bar charts (among others) to display metrics about phenomena using descriptive statistics including minimum, maximum, mean, sum, among others. Maps allow

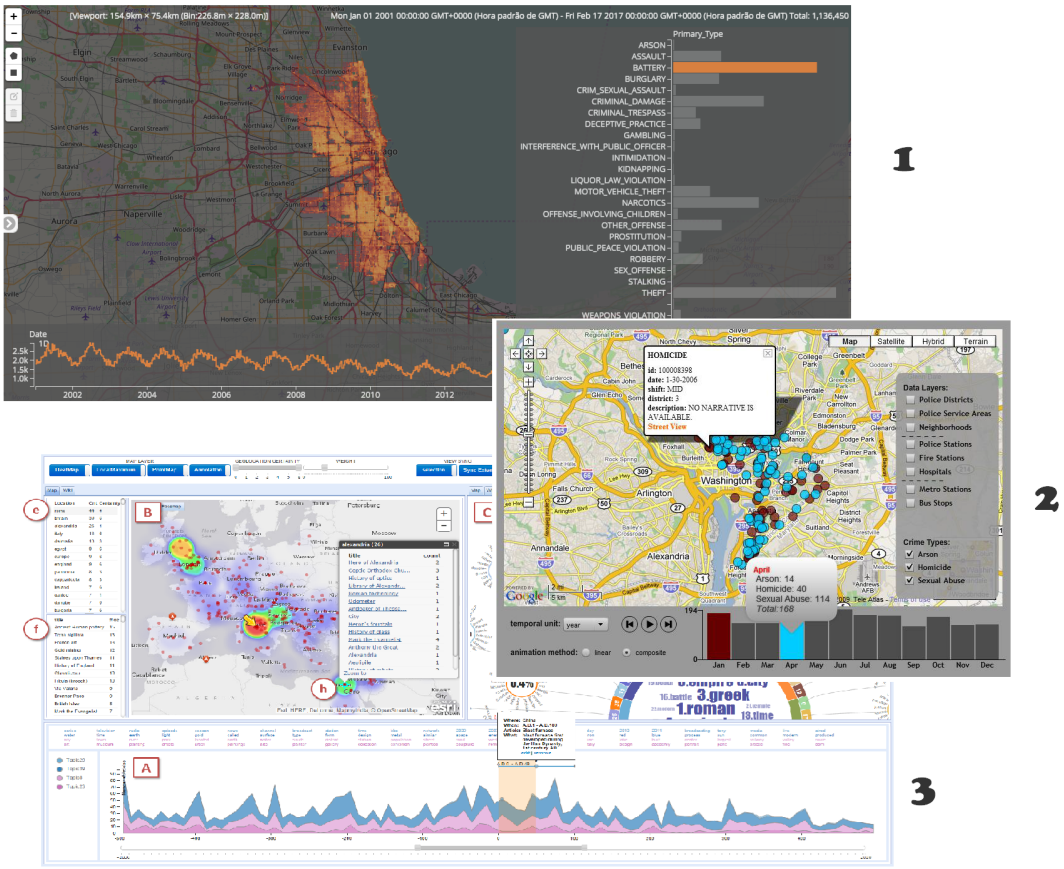


Figure 1.1: An overview of three approaches to explore spatiotemporal events: The tool *one* refers to (Lins et al. 2013); *two* refers to CrimeViz (Roth et al. 2010); and, *three* corresponds to VAiRoma (Cho et al. 2016).

us to understand how the intensity of a phenomenon is distributed throughout the geographic space, considering all or a particular time interval in which the phenomenon occurred; time series allow us to study how a phenomenon is distributed over time, concerning the entire geographic extent or a certain geographic region where a phenomenon occurs. Some examples of questions typically handled by these approaches are:

1. What is the spatial distribution of the phenomenon? Is it uniform or are there geographic regions with higher incidence?

2. How does the intensity of the phenomenon vary over time? Does it follow a decreasing or an increasing trend? Is there a cyclic pattern?

Let's assume that we need to explore the dataset of spatiotemporal events about crimes in the city of Chicago, in particular battery crimes, using the tool made available by (Lins et al. 2013) (see the tool *one* in Figure 1.1). Notice that, the time series (at the bottom) and the map are showing the number of crimes occurred. Using such a tool, or others that follow similar interfaces, we can perceive through the time series that the occurrence of

crimes has a cyclical behavior over time and follows a decreasing trend. Looking at the map, we can see that the phenomenon is almost uniformly spatially distributed.

Although this tool (or similar ones) provides a spatial and temporal overview of the phenomenon that is enough to answer several questions, namely those mentioned, there may be a number of other questions, which can be important for several organizations (Leipnik and Albert 2003; Ostfeld et al. 2005). Some examples are given:

1. Does the incidence of events follow some spatiotemporal pattern so that the events occur together close in time and space? In that case, are these spatiotemporal hotspots occurring across the entire geographic extent in which the phenomenon occurs, or do they arise only in some geographic regions?

2. Does the occurrence of events follow any contagious behavior [5]?

3. Do the events occur geographically dispersed over time? Or do they occur in a clustered way? Are there changes in the spatial distribution? Do the events sometimes occur in a dispersed form and sometimes they happen in clusters? Do these changes follow a particular pattern over time? Or does the phenomenon generally have a stable structure (let's say dispersed) and suddenly, in a particular moment in time, the events occur spatially clustered (moment of time as an outlier).

4. What is the pattern of occurrence of events over time concerning a particular administrative area? Do the events occur cyclically? Do the administrative regions close to each other following similar patterns of occurrence of events?

There is a substantial difference between the first set of questions presented and the second one. The first group is performing separate analyses of the spatial and the temporal dimension of the events, which are of limited value (Bogorny and Shekhar 2010; Møller and Ghorbani 2010; Wang and Yuan 2014). However, many pieces of information about the spatiotemporal dynamic of events like spatiotemporal patterns arise when one works with the spatial and temporal dimensions together, as the second group of questions requires, something that's challenging (Gabriel et al. 2013; Shekhar et al. 2015). One needs to account for the properties that distinguish spatiotemporal events from other types of data (Andrienko et al. 2010). These properties are dependency and heterogeneity (Yao 2003). Dependency can be explained through Tobler's first law: "*everything is related to everything else but nearby things are more related than distant things*" (Tobler 1970). The other property is the spatial heterogeneity and temporal non-stationarity, i.e., spatiotemporal events do not follow a similar distribution across the entire space and over all time. Instead, different geographical regions and temporal periods may follow different distributions.

[5] A "cloud" of events occur near in space and time that slowly changes its spatial location throughout time (Ostfeld et al. 2005).

Since, in general, present-day visual analytical approaches develop interactive visualizations to display the results of descriptive statistics, many patterns might not be captured (Kechadi et al. 2009; Miller and Han 2009; Shekhar et al. 2015). Even so, there are VA approaches (Maciejewski et al. 2010; Landesberger et al. 2012; Ferreira et al. 2013), usually developed to analyze a particular phenomenon (e.g., crimes), focusing on a particular kind of pattern (e.g, spatiotemporal hotspots). However, this may not be enough whether we aim at an approach independent from the application domain. Patterns might appear in many different forms and targeting a particular kind of pattern may leave many patterns to be detected. Nevertheless, the importance of patterns might depend on the specific application, the analysis question, and its concordance with domain knowledge (Keim et al. 2008; Sips et al. 2012).

1.1 The Level of Detail Matters

When one looks at spatiotemporal events, they can be expressed at different LoDs. The spatial location can be described using cells with different sizes (e.g., cells of 2 km^2 or 8 km^2), cities, counties or states; and the time can be specified with a detail of minutes, hours, months or years. The LoD reflects the size of the units in which phenomena are observed and often aggregated/summarized, most likely affecting our understanding of them (Marceau 1999; Andrienko et al. 2010; Laurini 2014).

A change in the LoD at which a phenomenon is observed can bring improvements to the analytical process (Camossi et al. 2008; Andrienko et al. 2010). From one LoD to another, some patterns can become easily perceived or different patterns may be detected. On one hand, different spatiotemporal phenomena exist and evolve at different LoDs, and on the other hand, a phenomenon may exhibit different patterns in different LoDs.

Some examples can be found in the literature. For example, Sips et al. 2012 describe a use case using glacial climate record data derived from an ice core from Dronning Maud Land, Antarctica. The ice core represents South Atlantic temperature in the past 150k years. Sips et al. 2012 provide the visualization method developed for scientists of the domain to detect strong temperature fluctuations. Those scientists report the highest fluctuations at the 10k year time scale (i.e., LoD) in comparison to other time scales. In that LoD, the highest fluctuations happened between the 10k-20k years before the present interval and 130k-140k years. This discovery allows them to make the hypothesis that the detected strong temperature fluctuations might be related to the 100k years cycle of the Milankovitch cycles[6]; Gabriel et al. 2013 investigate data about an epidemic in animals that occurred in 2001 at the Cumbria county in order to find out in what spatial and temporal LoDs the evidence of spatiotemporal hotspots emerge. In their approach, a change in the LoD means a change in the distance considered. The authors considered spatial distances of 5, 10, 15 km and temporal distances of 5, 10, 15 days. They found

[6]The 100k years cycle of the Milankovitch cycles describes the transition from a circular to an ellipsoidal orbit of the Earth around the Sun

evidence of spatiotemporal hotspots in temporal distances less than 10 days, and spatial distances less than 5 km. More examples can be found in the literature (Qi and Wu 1996; Wu et al. 2000; Dykes and Brunsdon 2007; Camossi et al. 2008; Plumejeaud et al. 2011; Goodwin et al. 2016; Zhang et al. 2016).

The LoD matters for the perception of phenomena and their underlying patterns, and often, there is no exclusive LoD to study phenomena (Dykes and Brunsdon 2007; Camossi et al. 2008; Andrienko et al. 2010; Sips et al. 2012; Goodwin et al. 2016). Although the LoD plays a crucial role in data analysis and pattern detection, this issue has been ignored with commercial analytical tools (e.g., *Qlik*, *Tableau Software*) and most of the state of the art proposals following a single LoD analysis approach (Dykes et al. 2005; Andrienko and Andrienko 2006; Power 2008; Zhang et al. 2012).

Users are left with the choice of the LoD(s) to look for patterns. The LoDs in which patterns can be perceived are often difficult to determine a priori (Sips et al. 2012). There might be several forms of patterns and these might be better perceived in some LoDs than in others, or even, different patterns might be perceived in different LoDs. If one considers n temporal LoDs and m spatial LoDs then there are $n * m$ spatiotemporal LoDs that can be studied in order to look for patterns as illustrated in Figure 1.2. Looking for patterns in different LoDs might be time-consuming and unproductive, following an analysis approache based on a single LoD (Camossi et al. 2008; Andrienko et al. 2010; Sips et al. 2012; Goodwin et al. 2016).

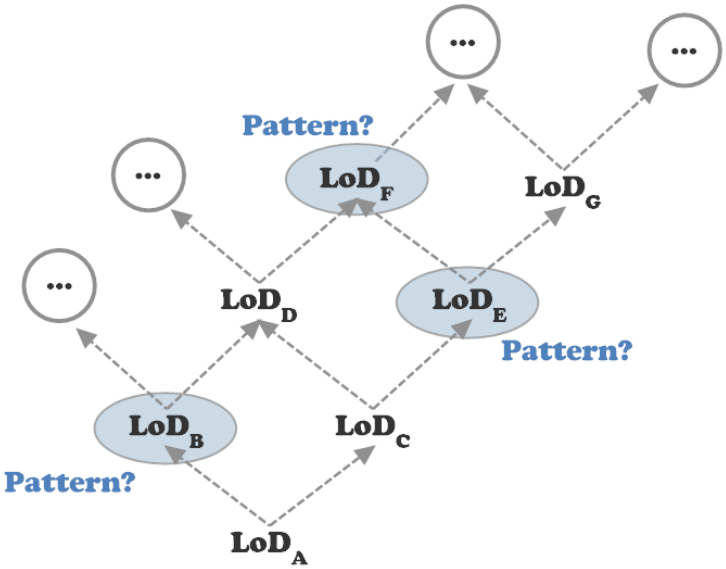


Figure 1.2: Many LoDs to look for patterns.

When someone is not familiar with a spatiotemporal phenomenon, i.e., an early stage of analysis, users can easily fall into a condition of information overload (Keim et al. 2008). By information overload, we mean, users face difficulties to develop a clear understanding of patterns that might be embedded in datasets of spatiotemporal events (Gabriel 2014; Robinson et al. 2016). From our point of view, this happens because VA

approaches communicate information mainly through descriptive statistics with some exceptions but focusing on a very particular pattern and application domain. Both options are not suitable for an early stage of analysis. Equally important, such approaches follow an analysis approach based on a single LoD, leaving the choice for the users. Without any help, this choice will remain a challenging task, as discussed, contributing for the information overload when conducting analyses over spatiotemporal events at early stages.

1.2 Research Statement

To face the information overload, from the user perspective, the VA area introduced an analysis approach: "*Analyze first, show the important, zoom, filter and analyze further, details on demand*" (Keim et al. 2008) known as the Visual Analytics Mantra (VA Mantra). It aims to provide the user with an understandable high-level overview of what is important, at an early stage of the analysis, thus reducing the data amount/complexity in order to make it analyzable and meaningful. Such approach aims to provide what is important from the available data to a given user at a given context so that he can conduct his analysis based on meaningful information.

When there is little information about a spatiotemporal phenomenon or the analytical goals are vague, i.e., an early stage of analysis, a user will probably experience an information overload. In spatiotemporal events, an approach to overcome this problem is to move from a single user-driven LoD to a multiple LoDs (simultaneously) analysis approach , providing the user with an understandable high-level overview of the underlying structure of the phenomenon for each LoD. By understandable high-level overview, we mean several hints about the distribution of events in space or/and in time that can provide a glimpse of the presence or absence of patterns. Following this approach, the user might detect very soon in what LoDs there are potential patterns and what kind of they are. According to his analytical goal and domain knowledge, the user would be able to better guide his analysis thus avoiding the information overload.

Despite the fact that literature recognizes the importance of LoD in the perception of phenomena and the need for users to study and explore phenomena across multiple LoDs (Dykes and Brunsdon 2007; Camossi et al. 2008; Andrienko et al. 2010; Sips et al. 2012; Goodwin et al. 2016), there are no approaches that work across several LoDs in the context of spatiotemporal events, and following the VA Mantra. The research problem addressed in this dissertation, can be stated as:

> *How can we help users explore phenomena logged as spatiotemporal events across multiple LoDs, simultaneously, helping them to understand in what LoDs there are patterns emerging?*

This broad formulation hides specific problems that cross different research areas

namely knowledge representation, data processing and visualization. In detail, this work seeks to address the following problems:

1. How do we enable representation and reasoning about spatiotemporal events at different LoDs? Making analyses across multiple LoDs requires modeling spatiotemporal events at different LoDs.

    a) What is a LoD? How do we formalize the concept of LoD?

    b) How do we model a phenomenon at different LoDs?

    c) Datasets of spatiotemporal events are collected at high LoDs. How do we follow a bottom-up automated approach in order to provide different phenomena's representations for each LoD?

2. With the datasets of spatiotemporal events available at multiple LoDs, we aim to provide analyses across them.

    a) How do we provide an understandable high-level overview about the underlying structure of the phenomenon for each LoD?

    b) How will the users inspect and compare the phenomenon perception across multiple LoDs?

    c) How do we provide an approach independent from the phenomenon without focusing on a particular analytical task or pattern?

A general overview of the problems solved and the results obtained during the research is given in the next section.

1.3 Research Goals and Contributions

The broad objective of the research introduced lies on enhancing exploratory analysis of spatiotemporal events, at early stages, by following analyses across multiple LoDs. Such approach aims to allow users to be able to inspect and compare the phenomenon's perception across multiple LoDs. To observe spatiotemporal events at different LoDs, we first need to represent and reason about spatiotemporal events at different LoDs.

1.3.1 Theory of Granularities

Granular computing has emerged as a paradigm of knowledge representation and processing (Yao et al. 2013), where granules are basic ingredients of information. Roughly, a granularity defines a division of a domain in a set of granules disjoint from each other. *Counties*, *States* are common examples of spatial granularities defined over the spatial domain; *Hours*, *Days* are common examples of temporal granularities. Erikson's stages

of psychosocial development (Erikson 1959) is an example of a granularity defined over the natural numbers [7].

Granules can be useful to express spatiotemporal events at different LoDs. Let's consider that we have a dataset of homicides events with the following structure: *homicide*(*S*, *T*, *Killer Age*). A homicide event originally logged as *homicide*((41.87803777, −87.62944228), 09/05/2015 20 : 42, 23) could be expressed at coarser LoD like *homicide*(*Illinois*, 09/05/2015 20*h*, *Early Adulthood*) using granules from the granularities *States*, *Hours*, *Erikson's stages*, correspondingly. However, in general, the granularity definitions found in the literature are applicable to particular domains like the time domain (Bettini et al. 2000) or the spatial domain (Camossi et al. 2006). This prevents us from representing and reasoning about events at different LoDs following a granular approach since there are different domains of reference underlying the events' features.

This PhD Thesis proposal introduces a formal Theory of Granularities (ToG) that allows the creation of granules over any domain of reference. This approach is more general than the current state of the art because the existing proposals appear as particular cases of the ToG proposed. Besides, it provides new instruments to reason over granules. Often, the domains of reference have relations defined between their elements. Four induced relations are proposed in order to transpose the relations defined in the domains of reference to the granules. Some of those relations have properties like symmetric, transitive, reflexive, antisymmetric, and antireflexive. The circumstances in which the induced relations inherit those properties were studied. This study goes together with formal proof conducted in the natural deduction system. These contributions led to one publication in the International Conference on Computational Science and Applications, and one publication in the International Journal Business Intelligence and Data Mining:

J. M. Pires, R. A. Silva, and M. Y. Santos, "Reasoning about Space and Time: Moving towards a Theory of Granularities," in Computational Science and Its Applications - ICCSA 2014, Springer, 2014, pp. 328–343

R. A. Silva, J. M. Pires, and M. Y. Santos, "A granularity theory for modelling spatio-temporal phenomena at multiple levels of detail," Int. J. Bus. Intell. Data Min., vol. 10, no. 1, p. 33, 2015.

1.3.2 Granularities-based Model

Using the ToG one can express individually spatiotemporal events at different LoDs. However, and up to this point, the concept of LoD is not defined and there is no model following an automated approach to generalize a phenomenon from one LoD to a coarser

[7] Erikson's stages: (i) Infancy - 0-1 years; (ii) Early childhood 2nd year; (iii) Preschool age 3–5 years; (iv) School age 6–12 years; (v) Adolescence 13–18 years; (vi) Early adulthood 19–39 years; (vii) Adulthood 40–64 years; (viii) Maturity 65-death

one. An automated approach is crucial as the number of LoDs from which one can perceive a phenomenon can be meaningful.

A granular computing approach was devised to model spatiotemporal phenomena at multiple LoDs labeled as the granularities-based model. This approach models a phenomenon through a collection of statements where, roughly speaking, granules are used in the statements' arguments. For example, $homicide(Illinois,\ 09/05/2015\ 20h,\ Early\ Adulthood)$ is an example of a statement concerning homicides in USA. Instead of single granules, complex descriptions can also be assigned to statements' arguments. For example, $tornado(RasterRegion(cells),\ Interval(09/05/2015\ 15:45,\ 09/05/2015\ 16:10),\ 20)$ stands for a tornado occurred on May 9th, 2015 between 15:45 and 16:10 pm, affecting a particular area with 20 victims. Complex descriptions are defined based on the general concept of granular term proposed in this PhD Thesis. Based on it, spatial granular terms (Cell and RasterRegion) and temporal granular terms (Instant and Interval) were formalized.

The granularities-based model defines the concept of LoD and follows an automated approach to generalize a phenomenon from one LoD to a coarser one. When changing a phenomenon's LoD a time interval can eventually be generalized to a time instant while a region might be simplified. This approach stands out from the related literature because (*i*) it models a phenomenon through statements rather than just using granules to model abstract real-world entities; (*ii*) as opposed to current granular computing approaches which are mainly concerned with indexing and aggregating data at different granularities, the granularities-based model provides different phenomena' representations for each LoD; finally (*iii*) a phenomenon can be expressed into other coarser LoDs in an automatic way. This research step led to two publications in the International Conference on Geographic Information Science:

> *R. A. Silva, J. M. Pires, M. Y. Santos, and R. Leal, "Aggregating Spatio-temporal Phenomena at Multiple Levels of Detail," in AGILE 2015, Springer International Publishing, 2015, pp. 291–308.*
>
> *R. A. Silva, J. M. Pires, M. Y. Santos, "When Granules are not enough in a Theory of Granularities," in AGILE 2017, Springer International Publishing, 2017, (in press)*

The granularities-based model was implemented in Java allowing to model phenomena stored as spatiotemporal events in a PostgreSQL database system. The module receives a dataset of spatiotemporal events as input and generalizes the phenomenon for each coarser LoD available.

1.3.3 SUITE Framework and Prototype

Through the granularities-based model, there is a phenomenon's representation for each LoD that leads us back to the research problem pursued in this dissertation. This PhD

Thesis proposal presents a framework for **SU**mmariz**I**ng spatio**T**emporal **E**vents (SUITE) across multiple LoDs. This framework builds summaries, at different LoDs, about phenomena logged as spatiotemporal events. Based on it, the users are able to inspect and compare the phenomenon's perception across multiple LoDs. As our framework does not make any assumption about the phenomenon and the analytical task, it can be widely used to get an overview of the phenomenon under analysis. The framework establishes five types of summaries working with space and time together. This allows us to frame and extend many proposals in the literature that create summaries of data in the proposed framework.

Using the SUITE's framework, one can have many summaries measuring different facets of the distribution of spatiotemporal events providing hints about the absence or presence of different kinds of patterns. To the best of our knowledge, there are no approaches that work across several spatial and temporal LoDs, and that are independent from the analytical task and the application domain in the context of spatiotemporal events.

To conduct analyzes in this new mindset, a web-based VA approach implementing the SUITE framework was developed. The prototype allows to visually inspect hints about the absence or presence of different kinds of spatiotemporal patterns at multiple LoDs, following a coordinated strategy among the visualizations provided. Moreover, one can study how patterns (spatial or non-spatial) evolve throughout time and also whether they happen only in some geographic regions or in all the geographic extent of the phenomenon. Notice that, these analyses occur always at multiple LoDs supporting them according to their analytical goals and domain knowledge in the choice of the suitable LoD to narrow their analyses in the future.

These research steps led to one publication in the International Conference on Geographic Information Science:

> *R. A. Silva, J. M. Pires, M. Y. Santos, and N. Datia, "Enhancing Exploratory Analysis by Summarizing Spatiotemporal Events Across Multiple Levels of Detail," in Geospatial Data in a Changing World, Springer International Publishing, 2016, pp. 219-238.*

A final remark about our contributions briefly introduced so far. The ToG is non-dependent of the data domain. The ToG is applicable independently whether the domain is natural or real numbers, discrete, dense, continuous or n-dimensional, for instance. The granularities-based model can be used to model any phenomena suitable to be modeled through statements and not necessarily just the ones logged as spatiotemporal events. In contrast, the SUITE's framework and prototype were developed not focusing in a particular application domain but just considering phenomena logged as spatiotemporal events. Nevertheless, the specificity of certain phenomena should not be ignored. Therefore, we allow that the set of summaries measuring different facets of the distribution of

spatiotemporal events providing hints about the absence or presence of different kinds of patterns be fine-tuned according to the phenomenon at study.

1.3.4 Evaluation

The evaluation of our proposals was conducted with two types of datasets of spatiotemporal events: (*i*) synthetic datasets; (*ii*) real datasets. In order to produce synthetic datasets a configurable generator of spatiotemporal events was used developed by (Gabriel et al. 2013) - R package (stpp). Using it, synthetic datasets with different spatiotemporal patterns like clustered, contagious, inhibitory, and infectious, in different spatiotemporal LoDs, with different cardinality were produced.

The real datasets used were: (*i*) forest fires in Portugal; (*ii*) the dataset made public by the Armed Conflict Location and Event Data Project[8] about conflict and protest data, occurring in Africa and Asia; (*iii*) crimes in the city of Chicago. These datasets contain information about different phenomena occurring in different spatial extents and different temporal extents.

Evaluation with users was considered but turned out to be challenging. We cannot consider any kind of user because the kind of analysis that we are aiming at is directed to domain experts of phenomena logged through spatiotemporal events. Thus, having a considerable number of users with balance gender, ages, backgrounds in order to allow a suitable evaluation is a real challenge.

That being said, the SUITE prototype was used to explore both types of datasets bearing in mind that when we explored synthetic datasets we knew beforehand in what spatiotemporal LoD a particular pattern was generated. For most of the datasets produced, the SUITE tool was able to provide a correct overview of the "phenomenon" allowing us to identify the LoD(s) in which the pattern generated occurs, and therefore, the LoDs that should be used to detail the analysis.

We then look for the patterns identified previously in the real datasets. Recognizing some of the patterns in the phenomena logged into the real datasets, in different spatiotemporal LoDs was easy. Afterward, we use the SUITE tool to explore the real datasets from several perspectives pursuing other forms of spatiotemporal patterns. Several patterns were identified at different spatiotemporal LoDs.

1.4 Thesis Structure

The remaining of the PhD Thesis is organized in the following structure:

Chapter 2 introduces fundamental concepts necessary to clearly understand the following chapters. It also presents the state of art theories, approaches and tools related to the matters addressed by the thesis.

[8] Website: `http://www.acleddata.com/`

Chapter 3 presents the theory of granularities that enables us to represent and reason about spatiotemporal events at different LoDs.

Chapter 4 presents the granularities-based model that allows us to model phenomena at different LoDs, following a bottom-up automated approach in order to provide different phenomena's representations for each LoD. Then, a demonstration case with a real dataset about tornadoes in USA is made.

Chapter 5 introduces a framework for **SU**mmariz**I**ng spatio**T**emporal **E**vents (SUITE) in order to help users explore phenomena logged as spatiotemporal events across multiple LoDs, simultaneously.

Chapter 6 presents the web-based VA prototype, called SUITE-VA, that implements our main contributions. Afterwards, the evaluation is presented in order to discuss if the broad objective was reached.

Chapter 7 concludes the PhD Thesis summarizing the results achieved and discussing several pointers for future work.

Chapter 2

Background and Related Work

The PhD Thesis addresses new foundations to model phenomena (particularly the ones logged through spatiotemporal events) at multiple LoDs so that we can have a bottom-up automated approach in order to provide different views of a phenomenon for each LoD. Also, it seeks to enhance an exploratory analysis of spatiotemporal events by following analyses across multiple LoDs. In this chapter, we present an overview of the fundamental concepts related to the thesis and discuss the state-of-art on its research field. This chapter is organized as follows.

A background about VA research area is given in Section 2.1. Since VA is a multidisciplinary field, an overview about the main research areas is provided. This section ends by presenting and discussing several VA approaches, supporting analyses based on a single LoD that have been developed to make exploratory analysis of spatiotemporal events.

To conduct analyses across multiple LoDs, we need to be able to represent and model data at different LoDs. Granular computing and its granularities concepts show potentialities to represent spatiotemporal events at different LoDs. For this reason, the state-of-art about granularities is discussed in Section 2.2. Furthermore, there are several works in the literature for modeling spatiotemporal phenomena at multiple LoDs. A discussion about the state-of-art of these approaches is given in Section 2.3.

Our broad goal is to enhance the exploratory analysis over spatiotemporal events through analyses at multiple LoDs. Therefore, the state-of-art about approaches conducting analyses at multiple LoDs is discussed in Section 2.4.

2.1 Visual Analytics

VA is the science of analytical reasoning supported by interactive visual interfaces (Thomas and Cook 2006). The VA Mantra is supported by automatic and visual analysis

methods with a tight coupling through human interaction in order to gain knowledge from data (Keim et al. 2008).

VA aims to combine human strengths (i.e., domain knowledge, cognitive capabilities) with the storage and processing capabilities of today's computers to gain insights into complex problems. Typical preprocessing tasks are like data cleaning, normalization, aggregation, or integration of heterogeneous data sources are performed. After that, an user may choose between applying visual or automatic analysis methods. If the latter is used first, data mining algorithms are used in order to compute patterns from data. Through the visualization methods, users are able to get insights from the generated patterns as well as interact with the automatic methods by modifying parameters or selecting other analysis algorithms.

Toggling between visual and automatic methods is a key characteristic of the VA (Keim et al. 2008; Silva et al. 2012). This is particularly useful when there is little information about the phenomena under study or the analytical goals are vague, once a user is directly involved in the analysis process and may adjust his analytical goals based on the results that he is getting. Phenomena recorded as spatiotemporal events fit this description, since frequently users aim to identify patterns that are unknown in advance. Equally important, the human involvement in the analytical process is crucial as the appropriate LoDs may depend on the specific application, the analysis in question, and the domain knowledge.

To deal with spatiotemporal data and the challenges that they it poses on data processing and interactive visualization methods, several research areas are involved in the visual analytics science, namely: (*i*) spatiotemporal data models that provide a formalism to represent and reason about the spatiotemporal data (Erwig and Schneider 2002; Galton 2009); (*ii*) knowledge discovery concerns to provide useful patterns dealing with the complexity of spatiotemporal data (Leung 2009; Mennis and Guo 2009; Miller and Han 2009); (*iii*) information visualization to develop novel visualization techniques in order to make them effective to visualize spatiotemporal data (Aigner et al. 2008; De Chiara et al. 2011). An overview of these research areas is given as follows.

2.1.1 Understanding Spatiotemporal Data

Everything that is spatial is also temporal. Spatiotemporal phenomena always occur at some location in some time period.

Time is generally modeled based on two temporal primitives: time instants or time intervals. Furthermore, *Time* is also modeled as linear or cyclic; continuous or discrete; with total order, partial order or branching (Frank 1998a; Aigner et al. 2008).

Time instants have no duration while a time interval I is the set of all time instants between a starting point (denoted by I^-) and an ending point I^+. Time points are limited to answering questions like whether two events took place at the same time or whether one event took place before the other. Similarly, time intervals are useful for answering

questions like whether the events started/ended together, whether events overlapped in time including all the questions that time points could answer. The theoretical approaches to modeling time have been studied in literature (Vilain 1982; Allen 1983; Frank 1998b).

There are three types of topological relations considering time instants and time intervals. Firstly, there are relations between pairs of time intervals. Such relations were defined in Allen's algebra (Allen 1983) which models all possible relative positions between two time intervals. There are 13 different possibilities. They are: after, before, meets, met by, during, contains, equal, finishes, finished by, starts, started by, overlaps, and overlapped by. Secondly, there are relations that can be held between time instants and time intervals or vice-versa. These relations were introduced by Vilain's algebra (Vilain 1982) which models all possible relative positions between a time instant and a time interval. There are 8 basic relations in the V-algebra: before, starts, started by, during, contains, finishes, finished by, after. Finally, there are relations that can be held between time instants. These relations come from point algebra: $<$, $\leq$,$=$,$\neq$,$>$,$\geq$.

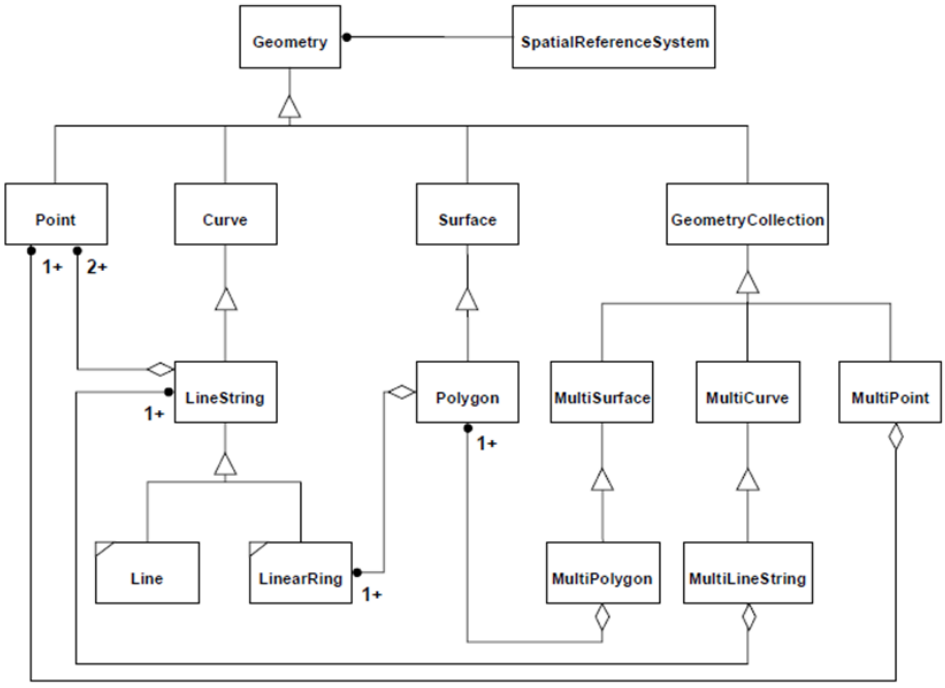


Figure 2.1: Geometry Class from OpenGIS specification from (Ryden 2005).

Space is represented by spatial data types. The OpenGIS Simple Feature Specification for SQL is an OGC specification that contains a norm defining spatial data types in a class diagram (Ryden 2005) (see Figure 1.1). From the abstract class Geometry derives Point, Line, Surface and GeometryCollection. Point is used to describe objects with zero dimensions, such as a traffic accident location. Line is used to represent objects like roads or rivers. Surface enables us to represent regions, areas or any other two-dimensional object such as counties, natural reservations, among others. GeometryCollection allows

for more complex objects resulting from the combination of multiple objects like a cluster of islands.

A spatial data type defines the properties and operations on objects in space. Operations on spatial data types include, for instance, the geometric intersection, union, and difference of spatial objects, the computation of the length of a line or the area of a region, the test whether two spatial objects overlap or meet, and whether one object is north or southeast of another object (Egenhofer and Sharma 1993; Schneider and Behr 2006).

Spatiotemporal data is being stored through different data types, capturing different spatiotemporal dynamics. Spatial data can be categorized into three models, i.e., the object model, the field model, and the spatial network model (Worboys and Duckham 2004). Spatiotemporal data, based on how temporal information is additionally modeled, can be categorized into three types, i.e., temporal snapshot model, temporal change model, and event or process model (Allen 1984; Kraak and Ormeling 2003; George et al. 2007; Yuan and Hornsby 2007; Alamri et al. 2014).

In the temporal snapshot model, spatial layers of the same theme are time-stamped. For example, if the spatial layers are points or multi-points, the set of temporal snapshots results into trajectories of points or georeferenced time series in case the variables are being observed at different times on fixed locations. Similarly, snapshots can represent trajectories of lines and polygons, or raster time series.

The temporal change model represents spatiotemporal data with a spatial layer at a given start time, and then, considers just incremental changes. For example, it can represent motion (i.e., speed and acceleration on spatial points) as well as rotation and deformation on lines and polygons.

Event and process models represent temporal information in terms of events or processes. The events are happenings (e.g., crime) whose properties do not change over time while the processes represent entities that are subject to change over time (e.g., movement of a person or car).

For different spatiotemporal data types different approaches have been proposed, in different research areas, so as to get more knowledge out of them. The research we conducted is focused on events has been already introduced. We discuss the current state of the art concerning the geovisualization and give an overview about automatic approaches. Subsequently several VA approaches mainly targeting spatiotemporal events were researched and further discussed.

2.1.2 Information Visualization Approaches

Information visualization is a broad research area which further divides into the following main categories: 2D visualization, 3D visualization, and color theory. 2D visualizations spans along 2 axes while 3D visualizations spans along 3 axes. Examples of standard 2D visualizations include bar charts, pie charts, line charts, maps among others. Concerning

3D visualizations, a well-known example is Google Earth[1]. Color theory deals with the suitable choice of colors in order to enhance the readability or help the visual analysis of data. For example, the work proposed by (Harrower and Brewer 2003) helps people to select good color schemes for maps and other graphics.

The number of visualization methods that have been developed is quite big (Bostock et al. 2011)[2]. We present some examples. Maps are essential to understand the location, extent and/or distribution of spatiotemporal events, spatial relationships, as detailed in (Bédard et al. 2007). The Parallel Coordinates (Inselberg and Dimsdale 1991) were designed to deal with multi-attribute data; the Circleview (Keim et al. 2004) method was designed to allow the user to observe temporal data in a cyclical way.

To understand the dynamic of spatiotemporal events, animated maps (Andrienko and Andrienko 2006) and change maps (Andrienko and Andrienko 2006) are often used. However, maps only represent multi-attribute data and dynamism (Bédard et al. 2007; Aigner et al. 2011); change maps are limited to small amounts of data and a few snapshots (each map representing a time instant or a time interval); the effectiveness of animated maps is therefore compromised (Tversky et al. 2002).

The role of visualization is an open issue when dealing with numerous spatiotemporal events at high LoDs. The visualization methods get easily cluttered and become difficult to analyze (Silva et al. 2012; Li et al. 2016). Visualization methods allowing the understanding of spatiotemporal events at different LoDs are still an issue that the information visualization's literature doesn't handle. This happens because a visualization needs to combine the spatial and temporal dimensions in a smart way in order be understandable, which, in our opinion and as you can see on the lines below, is quite challenging.

Aigner et al. 2011 make a comprehensive survey of techniques used for visualizing time-oriented data[3]. The visualizations were framed according to the following categories:

Frame of Reference: Abstract vs Spatial Abstract data (e.g., a bank account) has been collected in a non-spatial context and is not per se connected to some spatial location. Spatial data (e.g., spatial events) contains an inherent reference to spatial locations.

Number of variables: Univariate vs Multivariate Univariate data contains only one data value per temporal primitive, whereas in the case of multivariate (or multi-attribute) data each temporal primitive holds multiple data values.

Time Arrangement: Linear vs Cyclic Linear time corresponds to an ordered model of time, i.e., time proceeds from the past to the future. Cyclic time domains are composed of a finite set of recurring time elements (e.g., the seasons of the year).

[1] Google Earth: https://www.google.com/earth/
[2] D3 Gallery of Visualizations: https://github.com/d3/d3/wiki/Gallery
[3] The website: www.timeviz.net

Time Primitives: Instant vs Interval Time instants have no duration. A time interval, on the contrary, has a temporal extent greater than zero.

Visualization Mapping: Static vs Dynamic Static mapping maps time to space or maps time to visual variables (Bertin 1983) whereas dynamic mapping maps time to time. The former approach means that time and data are represented in a single coherent visual representation; as opposed to that, dynamic representations use the physical dimension of time to communicate the time dependency of the data.

Dimensionality: 2D vs 3D The representation of the temporal and spatial dimension in a visualization can be either two-dimensional or three-dimensional.

Along with the visualizations studied (Aigner et al. 2011), the site (`www.timeviz.net`) also allows users to filter the visualizations according to different categories proposed. By choosing just visualization methods developed for spatial data, keeping all the other categories, i.e., picking all the available visualization methods designed to analyze spatiotemporal data, the results gathered are shown on Figure 2.2.

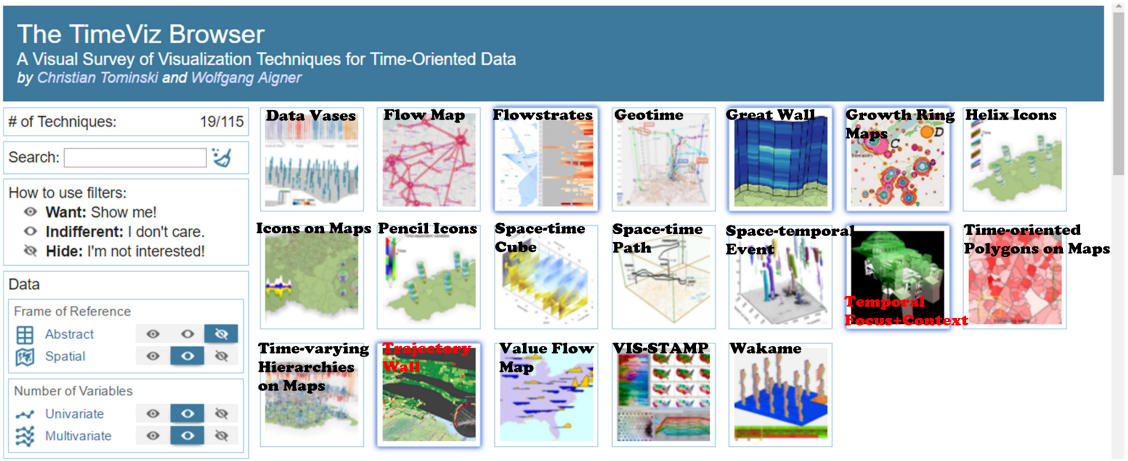


Figure 2.2: Screenshot showing the entire collection techniques to visualize spatiotemporal data, listed at `www.timeviz.net`.

From the 115 visualization methods surveyed by Aigner et al. 2011, just 19 were designed to display spatiotemporal data. From these 19, 4 (Flow Map, Flowstrates, Space-time Path, Trajectory Wall) were designed to show movements of objects over time, which is out of the scope of this work.

From the remaining 15, 4 (GeoTime (Kapler and Wright 2005), Space-time Cube (Kraak and Ormeling 2003), Time Varying-Hierarchies on Maps (Hadlak et al. 2010), Spatio-temporal event Visualization (Gatalsky et al. 2004)) make use of the space-time cube concept (X-Y to represent latitude and longitude and Z to represent time). In particular, Spatio-temporal event Visualization (Gatalsky et al. 2004) was designed specifically for displaying spatiotemporal events so that they are placed within the space-time cube

and the event's attributes can be encoded with visual variables like size, color, among others (Bertin 1983). However, 3D visualizations commonly suffer from occlusion and overplotting, making it difficult to grasp spatiotemporal patterns from their visual inspection.

A similar issue emerges from the 4 visualization methods (Data Vases (Thakur and Rhyne 2009), Helix Icons (Tominski et al. 2005), Pencil Icons (Tominski et al. 2005), Wakame (Forlines and Wittenburg 2010)) that use 3D diagrams over geographic regions as well as from the 2 visualization methods (Icons on Map (Fuchs and Schumann 2004), Value Flow Map (Andrienko and Andrienko 2004)) that use 2D diagrams to map the corresponding data values varying over time. Notice that, in order to use these last 6 mentioned visualization methods in a context of spatiotemporal events, we have to aggregate them by geographic regions. However, the diagrams can have a difficult readability if the number of geographic regions under study is high, or if they are quite close to each other.

The Time-oriented Polygons (Shanbhag and Rheingans 2005) might have also readability problems. This approach creates a partition of each polygon (2D) where each partition maps a value regarding a time period (using the color). The readability problems will emerge whether one considers small polygons or/and many time-periods. From the remaining results obtained, the most relevant for the analysis of spatiotemporal events might be: the Great Wall of Space-time (Tominski and Schulz 2012), VIS-STAMP (Guo et al. 2006) and Growth Ring Maps (Andrienko et al. 2011).

The Great Wall of Space-time (Tominski and Schulz 2012) creates a 3D wall based on a topological path over a cartographic representation. This wall is used to display how the data values associated to the geographic regions belonging to the path vary over time. This approach is not suitable to analyze spatiotemporal events because they are spread out in space and time. Therefore, we are not generally interested in a particular spatial path to analyze the phenomenon.

VIS-STAMP (Guo et al. 2006) is not a visualization method but a visual analytical approach that encompasses several visualization methods such as matrix plot, change maps, parallel coordinates that will be discussed later on in Section 2.1.4.

Growth Ring Maps (Andrienko et al. 2011) is a technique for visualizing the spatiotemporal distribution of events. Every spatiotemporal event is represented by one pixel. Each location (for example the centroid of spatial clusters of events) is taken as the center point for the computation of growth rings. The pixels (i.e., events) are placed around this center point in an orbital manner resulting in the so called Growth Ring representations. The pixels are sorted by the time at which the event occurred: the earlier an event happened, the closer to the central point the pixel is. Although this approach can be useful to provide a grasp on the spatiotemporal distribution of events, a clear understanding about when spatiotemporal hotspots occurred can be hard to achieve through visual inspection, for example. Furthermore, there might be others patterns that are not captured like changes in the structure of the spatial distribution of events throughout time.

As mentioned, the design of a visualization method that aims to combine the spatial and temporal dimension of data is not trivial. Perhaps that's why from 115 visualization methods surveyed by Aigner et al. 2011, we only have 19 visualization methods for spatiotemporal data. Their usage for spatiotemporal events was further discussed in this work, and in short, they have some problems handling spatiotemporal events. Another characteristic which is transversal to the visualization methods discussed is that they encode data into visual representations at certain LoDs. In fact, from our perspective, the visualizations should be used according to the LoD of the input data in spite of the issues identified for using them. For instance, Spatiotemporal event Visualization (Gatalsky et al. 2004) should be used when spatiotemporal events are provided at high LoDs (latitude and longitude coordinates) while the Time-oriented Polygons (Shanbhag and Rheingans 2005) should be used when the events are aggregated by some administrative level (e.g., counties) and by year.

In general, a visualization method produces a single representation of data. In order to make this representation effective, the visualization methods are designed taking into account the analytical goal and sometimes the data (Aigner et al. 2011). However, the analysis of spatiotemporal data frequently requires coordinated views in order to deal with the spatial, temporal, and thematic aspects of data simultaneously (Dykes et al. 2005). This approach has become standard in the recent applications of visual analysis because they directly support the expression of complex queries using simple interactions (Dykes et al. 2005; Scherr 2008; Weaver 2010).

In the project carried out by Lahouari et al. 2014, a set of geovisualization applications that allows us to study spatiotemporal phenomena in space and time were assessed. Such approaches were researched during a period of 6 months by 5 different people, independently, without having *a priori* a criterion of the target audience or the phenomenon or the technology implemented.

In total, 47 applications were studied and characterized based on ten criteria. Note that, an application might hold more than one value for a particular criterion. The first criterion considered was the types of spatial dynamics which the application was developed for. The ones considered as well as the percentage of applications studied for that type of spatial dynamic were: (*i*) spatiotemporal events - 25%; (*ii*) change in space (e.g., land use, Urbanization) - 38%; (*iii*) change of shape (e.g., black tide, cities' boundaries) - 6%; (iv) movements of individuals (e.g., daily trajectories of people) - 28%; (v) flux movement between places (e.g., home - work) - 19%.

Among the applications studied, 25% were developed to analyze phenomena logged as spatiotemporal events. From these, we left out from our next discussion the applications *Marine Traffic*, *Quick Route* and *ReRouteMe* as they are focused on the spatial movement. All the others are shown in Table 2.1.

The second criterion considered was the goal of the application. The goals considered are the simple presentation of data, the presentation of stories, the exploration and analysis, or the predictive analysis. From the 19% applications (the ones in Table 2.1), 44%

Table 2.1: Some applications evaluated by the project developed in (Lahouari et al. 2014)

	Application Goal	Space Representation	Time Representation	Spatial Granularity	Time Granularity	Source
CartoVista	Exploration	Map	Space	Multiple	Multiple	https://cartovista.com/demos/CustomFlash/CrimeAnalysisDemo/CrimeAnalysisDemo.html
CrimeViz	Exploration	Map	Space and Time	Multiple	Multiple	(Roth et al. 2010)
Data Rose - Ring Maps	Exploration	Map	Space	Simple	Multiple*	(Zhao et al. 2008)
HerbariaViz	Presentation	Map	Space	Simple	Multiple	https://www.geovista.psu.edu/herbaria/v3/index.html
How music travels	Story	Cartogram	Space and Time	Multiple*	Simple	http://www.thomson.co.uk/blog/wp-content/uploads/infographic/interactive-music-map/index.html
Mesure de la radioactivité dans l'environnement	Presentation	Map	Attribute	Simple	Simple	Not available
The Growth of Newspapers Across the U.S.	Presentation and Exploration	Map	Space	Simple	Simple	http://web.stanford.edu/group/ruralwest/cgi-bin/drupal/visualizations/us_newspapers
The Photographer's Ephemeris	Presentation and Prediction	Map	Space	Simple	Simple	http://photoephemeris.com/
Visualizing emancipation	Story and Exploration	Map	Space, Time, Attribute	Multiple	Simple	http://dsl.richmond.edu/emancipation/

make simple presentations of data (i.e., visualization of the location of events in space or/and in time) , 55% allow exploration of data, 11% make predictive analyses and 22% build stories with data.

The representation of space (i.e., location) was also defined as a criterion. A detailed discussion is not provided here as the majority of applications use maps to display the location of events.

Another criterion assessed was how time is represented. In this case, three different representations of time were considered: (*i*) through time as happens with animated visualizations (e.g., one second animation represents a year of time); (*ii*) using space like a line chart (e.g., 1 cm on a line chart represents a year time period); (*iii*) or as a data attribute (e.g., coloring events according to their date). This criterion is similar to the *Visualization Mapping* category discussed by Aigner et al. 2011. Most applications, 88% to be precise, use space to represent time through charts (e.g., Line Chart, Circleview) while 22% represent time as a data attribute and 33% use animation.

Another aspect evaluated was the ability of applications to view data in different temporal LoDs. The value for this criterion can be simple (the data is viewed in a single LoD and cannot be changed) or multiple (the data can be viewed in several temporal LoDs, one at a time or simulaneously). 56% of applications follow a simple approach while 44% follow a multiple approach. Similarly, in the case of space, 56% of applications follow a simple approach while 44% change the spatial LoD of analysis.

When the value for the previous criteria is multiple, the application might view data at multiple LoDs, one at a time, or simultaneously. In what concerns, only *Data Rose - Rings Maps* allows multiple temporal granularities simultaneously, while, in space, multiple spatial granularities are just supported by *How music travels*. Last but not least, none of the applications support data view at multiple spatial and temporal granularities (i.e., spatiotemporal LoDs) as it's aimed by this work.

The remaining criteria (Lahouari et al. 2014) are not detailed here as they do not have enough relevance for this work.

The information that results from this project provides us with some evidence. On one hand, the visualization methods discussed previously to display spatiotemporal data are not being adopted on spatiotemporal events probably because of the problems discussed. On the other hand, the applications allowing the analysis over spatiotemporal events are performing separate analyses of the spatial and the temporal dimensions (see Table 2.1), which are of limited value as pointed out in the beginning of the PhD Thesis. Therefore, patterns relating space or/and time may be hidden, and not identified, in the data that is usually displayed and analyzed. Furthermore, none of the approaches studied in this project, concerning spatiotemporal events, allows to view data at multiple spatial or/and temporal granularities, simultaneously, as it's aimed by this work.

2.1.3 Automated Approaches

Automated processing to extract knowledge from data can help users to handle the information overload. Spatial and spatiotemporal data mining studies the process of discovering interesting and unknown patterns that are potentially useful. Extracting patterns from spatiotemporal datasets is more difficult than extracting patterns from traditional alphanumeric data due to the complexity of spatiotemporal data (Shekhar et al. 2015). Such complexity comes from several challenges, starting by the properties like dependency and heterogeneity.

The dependency property can be explained through Tobler's first law (Tobler 1970): "*everything is related to everything else but nearby things are more related than distant things*" (Tobler 1970). For example, people with similar characteristics tend to cluster together in the same neighborhoods. Due to the spatial heterogeneity and temporal non-stationarity, spatiotemporal data does not follow an identical distribution across the entire space and over all time (Chawla et al. 2001; Miller and Han 2009). Instead, different geographical regions and temporal periods may have distinct distributions. Ignoring these properties may produce hypotheses or models that are inaccurate or inconsistent with the data set (Yao 2003; Miller and Han 2009; Wang and Yuan 2014; Shekhar et al. 2015). Furthermore, spatiotemporal datasets are embedded in continuous space and time, and thus many classical data mining techniques assuming discrete data (e.g., transactions in association rule mining) may not be effective (Shekhar et al. 2015). To handle such issues, spatial and spatiotemporal data mining algorithms have been proposed.

Spatial data mining is concerned with finding patterns in spatial data, ignoring the temporal dimension. The main output patterns are (Mennis and Guo 2009; Bogorny and Shekhar 2010): spatial association, spatial co-location, spatial clustering and spatial outlier. The spatial association rules represent a dependency relationship and take the follow form $X \rightarrow Y(c\%, s\%)$. Here, X and Y are two disjoint sets of items given a dataset, $c\%$ is the confidence (meaning $P(X|Y)$) and $s\%$ is the support (meaning $P(X \cup Y)$). The spatial association rule is an extension of typical association rules that considers the spatial properties and predicates in X and Y sets in addition to the attributes' values typically used. For instance, crimes occur frequently far from police stations. The spatial co-location rules represent subsets of features frequently located together like certain species of birds tend to use a certain type of trees as habitat. Spatial clustering is the process of grouping a set of spatial objects or events into clusters in such a way that objects or events in the same cluster have high similarity with each other, but are as dissimilar as possible to objects or events located in other clusters. An applicability of spatial clustering is to find hotspots of crime events, for instance. The spatial outliers represent observations which appear to be inconsistent with their neighborhoods. For instance, a store outperforms its neighbor competitors in sales numbers. Although spatial patterns do not take into account the temporal component, tracking the evolution of spatial patterns over time and detecting changes can be interesting.

Shekhar et al. 2015 provide a survey about spatiotemporal pattern families. The main families identified are spatiotemporal outliers, spatiotemporal coupling, spatiotemporal partitioning and summarization, and spatiotemporal hotspots.

A spatiotemporal outlier is a spatially and temporally referenced object or event whose non-spatiotemporal attribute values differ significantly from those of other objects in its spatiotemporal neighborhood. For example, spatiotemporal outlier detection can be used to detect the occurrence of unexpected events like crimes or traffic accidents. Spatiotemporal coupling patterns represent spatiotemporal objects or events which occur in close geographic and temporal proximity. For example, analysis of crime datasets may reveal frequent occurrence of misbehaviors and drunk driving after and near bar closings on weekends. Spatiotemporal clustering is the process of grouping similar spatiotemporal objects or events, and thus partitioning the underlying space and time. For example, partitioning and summarizing crime data, which is spatial and temporal in nature, helps law enforcement agencies find trends of crimes and effectively deploy their police resources (Chen et al. 2004; Malik et al. 2010). Spatiotemporal hotspots are regions jointly with certain time intervals where the number of objects or events is anomalously or unexpectedly high. For example, in epidemiology finding disease hotspots allows officials to detect an epidemic and allocate resources to limit its spread (Gabriel et al. 2013).

Several algorithms have been developed to compute spatial and spatiotemporal patterns and a survey on them can be found in (Roddick and Spiliopoulou 1999; Miller and Han 2009; Shekhar et al. 2015).

Often, the patterns have statistical expression. This way, spatial or spatiotemporal statistics are proposing quantitative analysis about the presence or absence of such patterns. The average nearest neighbor index (Ebdon 1985) (ANN) can give some hints about the presence of spatial clustering. If ANN's value is less than one, the pattern exhibits clustering. Otherwise the trend is toward dispersion. Getis-Ord General G (Getis 1992) measures how concentrated the high or low values are for a given study area. Positive scores indicate that the spatial distribution of high values is spatially clustered and the negative scores indicate that the spatial distribution of low values is spatially clustered. Getis-Ord General G measure might also suggest spatial outliers (Getis 1992). Global Moran's I (Moran 1950) measures the spatial autocorrelation or dependency based on feature locations and an associated attribute. When the spatial distribution of high values and/or low values in the phenomena is more spatially clustered than would be expected if underlying spatial processes were random, the Global Moran's I value will be positive. The spatiotemporal statistics methods like Knox (Knox and Bartlett 1964), Mantel (Mantel 1967) and the Jacquez k-nearest neighbor test (Jacquez 1996), measures the level of spatiotemporal interaction embedded in a phenomenon. More recently, Gabriel et al. 2013 proposed estimators to measure the spatiotemporal clustering/regularity in spatiotemporal point processes (equivalent terminology for spatiotemporal events with point as their spatial representation).

One challenge to mine spatiotemporal data results from the Modifiable area unit

problem (MAUP) (Openshaw and Openshaw 1984) or multi-scale (i.e., multiple LoD) effect since the results depend on a choice of appropriate spatial and temporal scales (i.e., LoDs) (Swedberg and Peuquet 2016). This means that patterns may be biased due to how data is aggregated/summarized. Analyses across multiple LoDs can make the MAUP identifiable or discarded sooner. For example, when a pattern is only visible in a specific LoD it can be further validated. One might conclude that the pattern suffers from MAUP and can be ignored or, if the phenomenon specifically operates there, it can be considered valid. Therefore, we argue that the analysis across multiple LoDs can attenuate the MAUP.

2.1.4 Visual Analytics Applications

There are several approaches to make analyses over data that emerge either from the academic or industry communities.

To the best of our knowledge, the more recent survey about commercial VA tools was done by (Zhang et al. 2012). About ten commercial VA tools (Tableau, Spotfire, QlikView, JMP (SAS), Jaspersoft, ADVIZOR, Solutions, Board, Centrifuge, Visual Analytics, and Visual Mining) were assessed by Zhang et al. 2012 namely in terms of automatic data analysis methods and visualization techniques implemented.

Regarding automatic data analysis methods, the authors divide the automated analysis functions implemented into statistics, data modeling, and data projection. The first category includes statistics functions for: (*i*) univariate analysis that operate on one dimensional data, for example, the calculation of the mean, minimum and maximum, and standard deviation; (*ii*) bivariate analysis that reveals correlation of two variables, for example, Pearson correlation coefficient; and (*iii*) multivariate analysis that models the relations over multiple dimensions. From the systems studied, all provide some simple statistics methods for univariate and bivariate analysis, but multivariate analysis is only supported by QlikView, Spotfire, JMP and ADVIZOR.

The data modeling category aims to find patterns using various data mining algorithms. The most commonly implemented algorithms include clustering algorithms, classification or network modeling. Among the systems studied, Spotfire, JMP, Centrifuge implement partitioned based and hierarchical clustering. To the best of our knowledge, the clustering algorithms are applied to non-spatial and non-temporal data.

The third category (data projection) describes dimension reduction techniques that can be applied to transform high dimensional data into lower dimensional space. Such transformation leverages the dimensionality problem by reducing the number of dimensions prior to analysis or visualization while keeping the essence of the data intact. The result is often used to generate 2D or 3D projections (typically scatter plots) of the data. The commonly used dimension reduction techniques are Principle Component Analysis (PCA), Multidimensional Scaling (MDS) and Self Organizing Map (SOM). In this PhD Thesis, we aim to work with space and time together in order to understand spatiotemporal patterns in datasets of spatiotemporal events. Dimensionality reduction is supported

by QlikView, Spotfire, and JMP. The previous discussion was summarized by the authors in Figure 2.3.

	Statistics			Data Modelling				Data Projection		
	Univariate	Bivariate	Multivariate	Clustering	Classification	Network Modelling	Predictive Analysis	PCA	MDS	SOM
Tableau	✓	✓	-	-	-	-	-	-	-	-
QlikView	✓	✓	(✓)*	(✓)*	(✓)*	-	(✓)*	(✓)*	(✓)*	(✓)*
Spotfire	✓	✓	(✓)*	P / H	(DT, NB, ANN)*	-	*(AR,HW)**	(✓)*	(✓)*	-
JMP	✓	✓	✓	P / H	DT, ANN	-	✓	✓	-	✓
Jaspersoft	✓	✓	-	-	-	-	-	-	-	-
ADVIZOR	✓	✓	✓	-	SVM		MVLR	-	-	-
Visual Analytics	✓	✓	-	P / H	-	✓	-	-	-	-
Centrifuge	✓	✓	-	P / H	-	✓	-	-	-	-
Visual Mining	✓	✓	-	-	-	-	-	-	-	-
Board	✓	✓	-	-	-	-	-	-	-	-

(...): only with additional upgrades*
DT, NB, ANN: decision tree, naive bayes, artificial neural network
SVM: support vector machine
P / H: partitioned based clustering / hierarchical clustering
AR, HW: ARIMA, Holt-Winters
MVLR: multivariate linear regression

Figure 2.3: Summary about Automatic Analysis Methods supported by the commercial VA tools (Zhang et al. 2012).

Concerning the visualization, the authors classify the visualization techniques by the type of data, namely (*i*) numerical data; (*ii*) georelated data. The authors conclude that the number of visualization techniques that are implemented by the surveyed VA systems is small when compared to the number of techniques that are available from research (Bostock et al. 2011). In other words, in general, the methods implemented for numerical data are Histograms, Scatter plot, Heatmap, Paralell Coordinates, Scatterplot Matrix; and for geo-related data there are just maps (see Table 2.4).

	Numerical Data					Geo-related Data
	Bar- line-pie- Chart Histogram	Scatterplot	Heatmaps	Parallel Coordinates	Scatterplot Matrix	Projection on Map
Tableau *	✓	✓	✓	(✓)	✓	✓
QlikView	✓	✓	(✓)	(✓)	✓	(✓)
Spotfire	✓	✓	✓	✓	✓	✓
JMP *	✓	✓	✓	✓	✓	✓
Jaspersoft	✓	✓	✓	-	-	✓
ADVIZOR	✓	✓	✓	✓	✓	✓
Visual Analytics	✓	✓	✓	-	-	✓
Centrifuge	✓	✓	✓	-	-	✓
Visual Mining	✓	✓	✓	(✓)	-	✓
Board	✓	✓	✓	(✓)	-	✓

(✓): not available as default, user interaction required (eg., transform line charts to parallel coordinates)
** tool that suggests appropiate visualizations to the user*

Figure 2.4: Summary about Visualization techniques supported by the commercial VA tools (Zhang et al. 2012).

Last but not least, and although the authors have not discussed, these tools follow an analysis approach based on a single LoD, which is chosen by users. Furthermore, they are general purpose VA tools (Stewart et al. 2015) and they are not designed to handle datasets of spatiotemporal events specifically. Therefore, spatial or spatiotemporal patterns will be difficult to be identified using the commercial VA tools studied by Zhang et al. 2012 as they just support separate analyses of the spatial and temporal dimension

of data.

So far, we gave an overview of VA commercial tools, discussed the main visualization techniques used for spatiotemporal data, and the main output patterns that are searched in them. As we aim to an approach that enhances exploratory analysis of spatiotemporal events, a detailed discussion about VA approaches found in the literature that analyses spatiotemporal events based on a single LoD is provided below.

Guo et al. 2006 developed a visual analytics software package called, VIS-STAMP, that couples computational, visual, and cartographic methods for exploring and understanding spatiotemporal and multivariate data. It can help analysts investigate complex patterns across multivariate, spatial, and temporal dimensions via clustering, sorting, and visualization.

The input data is a space-time-attribute cube as illustrated in Figure 2.5. Each cell in this cube is defined by a specific spatial object (e.g., Texas), a specific time (e.g., year 2000), and a specific variable (e.g., sales percentage for the energy industry). A time-attribute slice (see Figure 2.5b) can be seen as a series of multivariate profiles (one for each year - Figure 2.5c), or a set of time series (one for each variable - Figure 2.5d).

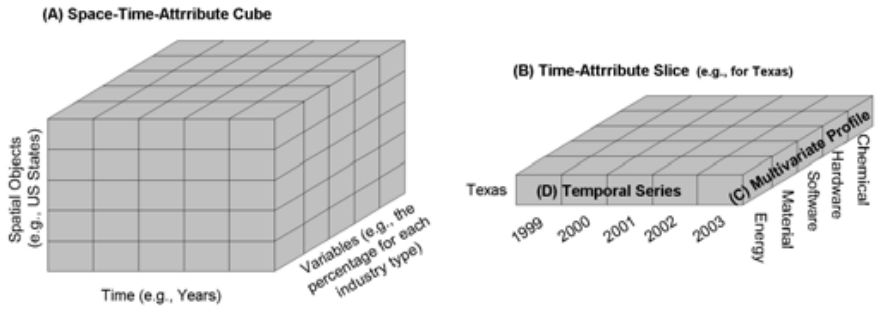


Figure 2.5: Illustration of the space-time-attribute cube(Guo et al. 2006).

An overview of the VIS-STAMP interface can be seen in Figure 2.6. The core of the system lies on the self-organizing map (SOM) that is used for multivariate clustering, sorting, and coloring. SOM takes a set of temporal series or multivariate profiles as input. The clusters computed are displayed at bottom-right and the circle' size is proportional to the number of data items it contains. The parallel coordinates (PC) (bottom-left) are used to display the clusters identified. If the SOM input is a set of temporal series, the coordinates of PC will be the time's values. In case the input is a set of multivariate profiles, the coordinates of PC will be the attribute's values. Accordingly, the matrix view's columns (top-left) represent attribute's values and its rows stand for geographic regions. The matrix view's columns represent the time's values. The map view follows the change maps approach and shows choropleths, one for each time point or attribute value, like in the matrix view. The color scheme is consistent across all views.

In Figure 2.6, the SOM was used over temporal series based on the space-time-attribute cube displayed in Figure 2.5 with real data. Using this approach, one can

Figure 2.6: An overview of the VIS-STAMP interface (Guo et al. 2006).

study the variation of temporal patterns across geography and multiple categories (e.g., industry types). The colors represent similar temporal trends as displayed by the PC. For example, red and dark red colors represent a most recent growth in sales displaying low sales for most years but rapid rises in 2003 concerning the nonprimary high-tech (NON) industry as can be seen in the matrix view.

Although this approach allows the search for spatiotemporal patterns, this can only be done for one spatiotemporal LoD at a time. In the example mentioned, the underlying spatiotemporal LoD was *State*, *Year*. Besides, if one intends to use this approach to look for spatiotemporal patterns in events, we need, beforehand, to aggregate them for coarser spatial and temporal LoDs. This actually happened in a later work they proposed (Guo and Wu 2013). Even so, if those LoDs are not coarser at all, we might end up with too many geographic regions and time periods, which will probably cause readability problems in the matrix and spatial view, and therefore, the analysis will turn out to be difficult.

Maciejewski et al. 2010 propose an approach to identify spatiotemporal hotspots of

events like crime events or syndromic ones. An overview of its interface is displayed in Figure 2.7. The main viewing area is the map, and the three charts on the right that allow users to view a variety of data sources simultaneously for a quick comparison of trends across varying hospitals/precincts or data aggregated over spatial regions. These views are synchronized. Additionally, both the map and the time series are linked to the time slider at the left side of the screen. This allows users to view the spatial changes in the data as they scroll across time. Furthermore, temporal controls are also employed on the left side denoted as "aggregate" and "increment". The aggregate function allows the user to show all data over a period of α days. The increment function allows the user to step through the data by increments of 1, 2, 3, ... days. On the map view, one can choose to look at the events' spatial locations, the events' spatial locations grouped based on nearest-neighbors, the events aggregated by some administrative level or to use the geospatial heatmap proposed as illustrated in Figure 2.7. When the heatmap is used, the percentage of events over the total events occurred on the period of time chosen in the time slider is displayed.

The authors developed an analytical approach focused on finding spatiotemporal hotspots. Furthermore, in order to understand in what spatial and temporal LoDs the spatiotemporal hotspots emerge, or how they are better perceived, the users might have to try several levels of aggregation (i.e., LoDs). Using the map, events can be displayed at different LoDs like, for instance, the actual spatial locations, aggregated based on their neighbors, or even, aggregated by administrative levels. On the other hand, the data might be aggregated using different ranges of days (e.g., one, two, three) according to the "aggregate" control in the interface. Bearing this in mind, finding the suitable spatiotemporal LoD(s) to explore the data can be challenging.

As previously presented, Lins et al. 2013 propose a compressed hierarchical data structure in order to hold huge amounts of spatiotemporal events in memory. In addition, the authors implemented some web-based applications to explore real datasets of spatiotemporal events. They developed an application to perform analysis over crimes occurred in the City of Chicago as shown in Figure 2.8. Other applications can be tested in their website[4].

In this case, the interface is composed by a map showing the location of events. The spatial LoD at which the events are displayed changes according to the zoom level. However, the same behavior was not registered when the time series was analyzed. Besides that, the interface contains a line chart with the number of events aggregated by day. This approach does not focus on a particular analytical goal but these are addressed using the descriptive statistic COUNT. Another characteristic is the fact that this approach is independent from the phenomenon. Furthermore, and although they have spatiotemporal events available at different spatial and temporal LoDs, their analyses are conducted using one spatial or temporal LoD at a time, separately.

[4] http://www.nanocubes.net/

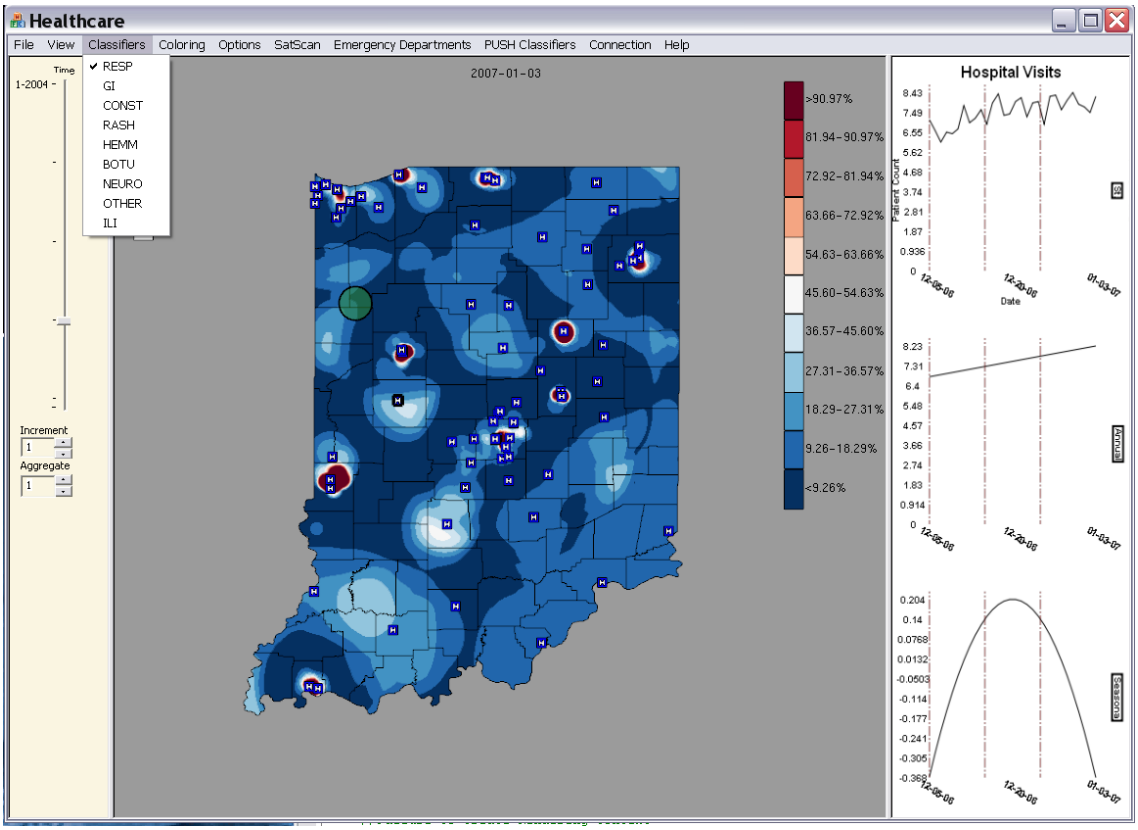


Figure 2.7: An overview of the VA system proposed by (Maciejewski et al. 2010).

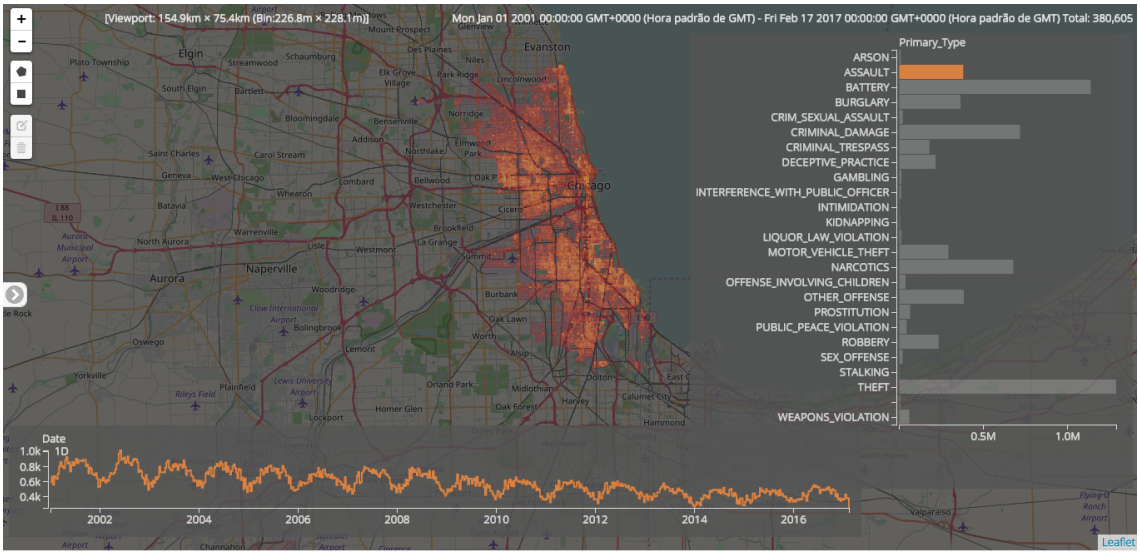


Figure 2.8: An overview of an application developed by (Lins et al. 2013).

Ferreira et al. 2013 develop a visual environment to explore taxi trips, called *TaxiVis*. Analyzing it, the input data are events of taxi pickups and taxi drop-offs that happened in New York City. An overview of the tool's interface is displayed in Figure 2.9a. Users can specify queries over all the dimensions of the data and explore the attributes associated with the taxi trips (e.g., What are the geographic regions with higher demand for taxis?). Besides standard analytics queries, *TaxiVis* supports origin-destination queries that enable the study of mobility across the city (e.g., What is the average trip time from Midtown to the airports during weekdays?). This kind of queries that aim to understand movement patterns are not the goal of this PhD thesis. Another important feature of the system is the ability to compare spatiotemporal slices through multiple coordinated views. Users can interactively compose and refine queries changing the queries parameters like the attribute to be analyzed, the spatial LoD, the temporal LoD of aggregation, among others. For example, the spatial LoD might be based on neighborhood, administrative level, or even, user-defined geographic region whereas the temporal LoD can be the hour, the week, the month, the year.

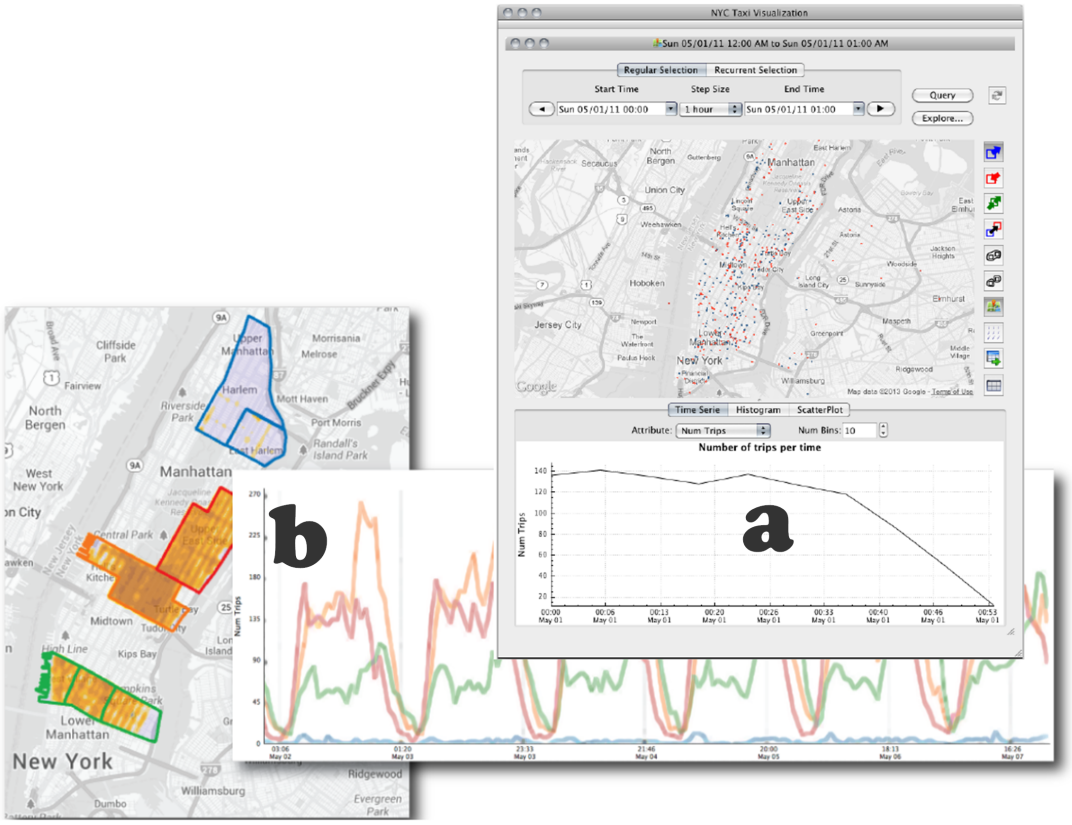


Figure 2.9: An overview of the VA system proposed by (Ferreira et al. 2013).

This approach supports exploratory analysis about taxi pickups and taxi drop-offs without any particular analytical task in mind. They are addressed using descriptive statistics that result from the separate analysis of the spatial and the temporal dimension

of data. For example, Figure 2.9b, shows how the number of pickups vary over the first week of May 2011 in four different geographic regions. Furthermore, the spatiotemporal LoD at which the data is explored is visually driven by the user according to his analytical goals.

Another VA approach developed was VAiRoma (Cho et al. 2016). Although this approach was not developed particularly to analyze spatiotemporal events, the type of data handled has some similarities. From the entire collection of English Wikipedia articles, the authors extract the ones related to the Roman History based on key words like "Rome","Roma", and "Roman". Afterwards, they preprocess the articles in order to place them in space and time based on their content. In the end, they manage to have a dataset about the Roman history where facts associated to articles are referenced in space and time.

An overview of the VAiRoma's interface is given in Figure 2.10. The interface is composed of three main views: geographic, timeline and topic. The timeline view (Figure 2.10A) presents temporal topical trends of the Wikipedia collection over 4000 years (2000 BC to 2010 AD). Each point in the timeline is representing the number of articles related to a certain topic(s). The map (Figure 2.10B) shows the location regarding the facts described in the articles selected via the time period, or by topic which can be done using the topic view (Figure 2.10C).

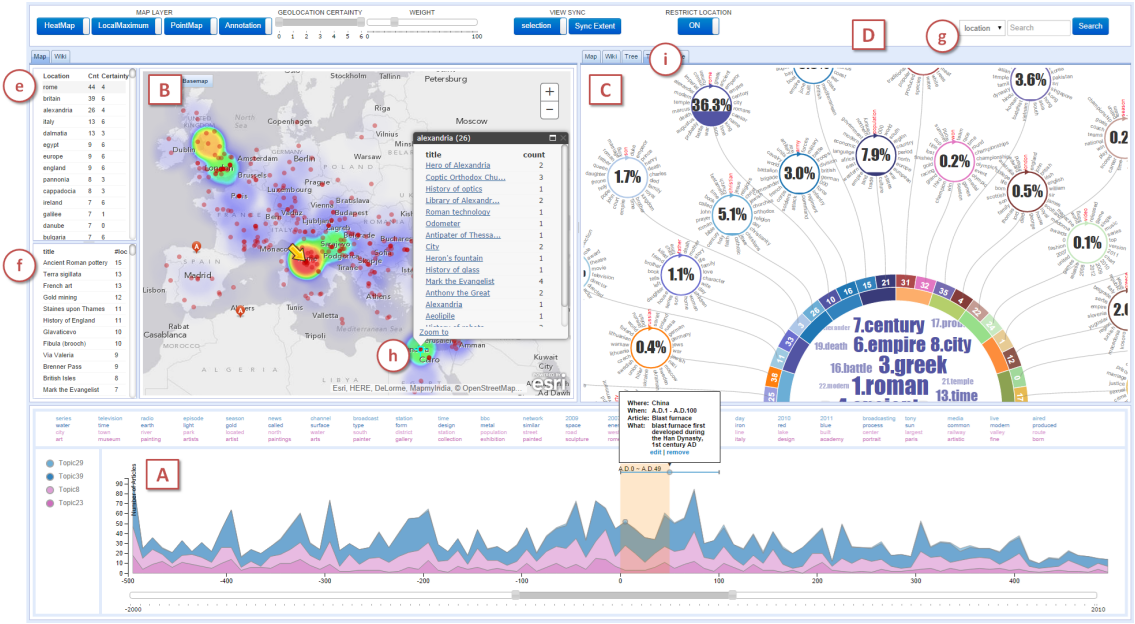


Figure 2.10: An overview of the VAiRoma system proposed by (Cho et al. 2016).

VAiRoma focuses on constructing a narrative of the whole Roman history from ancient times, through the Empire, to modern times, and not on extracting spatiotemporal patterns that might have happened.

Some of the VA approaches discussed support separate analyses of space and time and these analyses are performed at one spatiotemporal LoD at a time like the works

(Ferreira et al. 2013; Lins et al. 2013; Cho et al. 2016) discussed in detail here. More similar approaches were found (Kisilevich et al. 2010; Malik et al. 2010; MacEachren et al. 2011; Andrienko et al. 2013).

Others approaches support analyses that look for spatiotemporal patterns like Guo et al. 2006 or Maciejewski et al. 2010. However, these kinds of approaches follow analyses based on a single LoD, and in some cases, they are developed for the detection and exploration of a particular spatiotemporal pattern in a particular domain application (Chae et al. 2012; Thom et al. 2012; Wang et al. 2013). As opposed to that, we aim to give an overview of the presence of absence of spatiotemporal patterns at different LoDs simultaneously without focusing in a particular application domain but just considering phenomena logged as spatiotemporal events.

2.2 Granular Knowledge Representation

Granular computing has emerged as a paradigm of knowledge representation and processing, where granules are basic ingredients of information. Granular computing is an umbrella term to cover any theories, methodologies, techniques, and tools that make use of granules in complex problem solving (Yao et al. 2013). In granular computing, there are several formal platforms in which information granules are defined and processed wherein some approaches are defined based on set theory and others on top of Fuzzy sets, Shadowed sets and Rough sets (Bargiela and Pedrycz 2012). This work focused on approaches defined based on set theory, once we aim to model data at different LoDs, and not to model aspects of uncertainty or imprecision in the concepts. Since the granularities definitions found in the literature were developed for specific data domains, we end up proposing a Theory of Granularities (ToG) applicable to any data domain (space, time, and other attributes). Furthermore, it gave us the foundations to support an automated approach for data generalization for coarser LoDs.

The granularities definitions proposed in the literature are focused mainly in temporal or spatial domains. A temporal granularity, proposed by Bettini et al. 2000, is a sequence of temporal granules, each one composed by a set of time instants. For example, December 2016 can be a temporal granule. Consider a time domain T as a set of totally ordered time instants. A temporal granularity G_t is a mapping from an index set (e.g., the natural numbers) to subsets of the time domain. Suppose that i, k and j are elements of an index set. A temporal granularity needs to satisfy the following conditions:

- if $i < j$ and $g_t(i)$ and $g_t(j)$ are non-empty, then each element in $g_t(i)$ is less than all the elements in $g_t(j)$;

- if $i < k < j$ and $g_t(i)$ and $g_t(j)$ are non-empty, then $g_t(k)$ is non-empty. Each non-empty $g_t(i)$ in the above definition is called a temporal granule.

These conditions impose the following: temporal granules of the same temporal granularity cannot overlap and non-empty temporal granules must preserve the order given by the index set. Moreover, we cannot have an element (from the index set) mapped to the empty set between any two elements mapped to non-empty subsets. Accordingly, $Weeks$, $Years$ are examples of temporal granularities. Notice that, in the logic community, an independent line of research on representation and reasoning with multiple granularities investigated classical and non-classical logic extensions based on multi-layered time domains. More details about this approach can be found in (Euzenat and Montanari 2005).

A spatial granularity G_s is a set of spatial granules, each one being a portion of a spatial domain. Camossi et al. 2006 define spatial granularity as a mapping from an index set to subsets of the spatial domain (assumed as 2-dimensional) such that: if $i \neq j$, and $g_s(i)$ and $g_s(j)$ are non-empty then $g_s(i)$ and $g_s(j)$ are disjoint. No order is required among the spatial granules, but two spatial granules of the same granularity cannot overlap. Examples of spatial granularities are: $Countries$, $Cities$, among others. The spatial granularity definition is further extended by Belussi et al. 2009 in order to also represent the relations between spatial granules (e.g., direction-based relations, distance-based relations).

The proposals regarding spatial granularities discussed so far are focused on vector data. As opposed, Pozzani and Zimányi 2012 propose a framework focused on raster data. The authors define a spatial granularity σ as a total function from two-dimensional coordinates in Z^2 to a label set L such that $\sigma : Z^2 \rightarrow L$. This way, given a cell $c \in Z^2$, $\sigma(c)$ represents the label associated to c. Unlike the previous approaches, a granule corresponds to the sets of all cells sharing the same label.

Either on vector-based granularities or raster-based granularities there are proposals to handle the evolution of spatial granularities. Under the terminology used in the literature, sometimes these correspond to spatiotemporal granularities (Belussi et al. 2009; Pozzani and Zimányi 2012), which from our point of view is not the most accurate term to be used. We would reserve the term spatiotemporal granularity to mention granularities where each granule refers to a portion of a R^3 (e.g, if we assume the space as R^2 and time as an additional dimension). In any case, the evolution of spatial granularities is necessary to handle changes of spatial granularities over time. For example, a country's administrative division may change over time. The evolution of granularities is crucial to handle such scenarios.

Belussi et al. 2009 propose a definition for handling the evolution of spatial granularities. It has two components $< tG, E >$. tG is a temporal granularity and E is a mapping that to each time moment t, bound by a lower and a upper bound, associates the spatial granularity valid on it. Regarding the work by Pozzani and Zimányi 2012, it follows the previous approach applied to their spatial granularity definition. Another approach found in the literature introduces the concept of spatiotemporal granularity (Wang and Liu 2004).

Zadeh considers granular computing as a basis for computing with words (Zadeh 1998). Granularities should give us "words" (i.e., granules) to make statements about phenomena. The granule concept should be applicable to any domain of reference and not necessarily just to the spatial or temporal domain.

Keet 2008 shares this mindset and developed a formal, domain-independent theory of granularity that can be used for computational reasoning. This theory was developed to model phenomena at different LoDs and applied it to biological sciences. A domain of reference can be granulated with a certain criterion following a type of granularity that defines a granular perspective, which in turn contains granular levels. In her comprehensive theory, there is a proposal for a taxonomy of types of granularity. This taxonomy makes explicit both the ways of granulation, and how entities are organized within a granular level. For example, one type of granularity is denoted by nrG: *levels of non-scale-dependent granularity are ordered according to one type of relation in a perspective (e.g., structural: part of, spatially: contained in)*. For example, administrative divisions, the granulation criterion, can be used to define a granular perspective containing granular levels of type nrG (considering the spatial relation contained in). This granular perspective can be composed by three granular levels, for instance: *Countries*, *States*, *Municipalities*. In this case, the granular level and the spatial granularity proposed in (Camossi et al. 2006) have similar interpretations. Note that, the concept of granular perspective make explicit the characteristics of hierarchies of granularities, something that's left implicit throughout the literature on granularity (Keet 2008). However, granular levels are static in the sense that they don't handle a temporal evolution. Furthermore, Keet's theory of granularities has no full support for dealing with the complexity of temporal granularities (Keet 2008) which is crucial for modeling spatiotemporal phenomena.

Bravo and Rodríguez 2014 also make a generalization of the concept of granularity for any domain through the concept of *domain schema*. However, this work does not provide a comprehensive theory of granularities like Keet 2008 but rather their work take another direction that will be discussed in Section 2.3.3.

Granular computing shows itself useful to model phenomena at different LoDs because granularities can be related through relationships allowing one to compare and relate granules belonging to different granularities (Bettini et al. 2000; Camossi et al. 2006). Two commonly used relationships between granularities (spatial or temporal) are given. A granularity G groups into H if each granule of H is equal to the union of a set of granules of G. For example, *Days* groups into *Weeks*, but *Weeks* do not group into *Months*. A granularity G is finer than H if each granule of G is contained in one granule of H. For instance, *Portugal's parishes* is finer than *Portugal's districts* but *Rivers* is not finer than *Countries*. Some relationships are only applicable to some kind of granularities. For instance, in temporal granularities, we found groups periodically into or shift equivalent relationships (Bettini et al. 2000). More details about granularities' relationships can be found in (Bettini et al. 2000; Belussi et al. 2009; Pozzani and Zimányi 2012). Additionally, we can perform operations over granularities. In general, the operations are

proposed to automate the creation of new granularities. More details about this subject can be found in (Bettini et al. 2000; Keet 2008; Belussi et al. 2009).

Granules, granularities and the relationships between them are fundamental concepts to understand granular computing approaches that model phenomena at different LoDs. Despite the several proposals for granularities, the concept is being narrowed to a division of a domain in a set of granules disjoint from each other. However, to the best of our knowledge, we do not find on the literature a theory of granularity that enables us to define granularities over any domain, to reason about granules with known relations of the domain of reference and to handle the evolution of a granularity defined over any domain.

2.3 Modeling Phenomena at Multiple Levels of Detail

To model spatiotemporal phenomena at multiple LoDs, spatiotemporal models have been investigated, proposed by different researchers in different research areas like multirepresentation, multiresolution, granular computing, and compressed data structures. A discussion of them is provided.

2.3.1 Multirepresentation Approaches

Multirepresentation provides different points of view from a spatiotemporal phenomenon allowing the observation of the same geographical space and/or interval of time, from different perspectives. For example, we can have a representation of a country in terms of unemployment and another representation of the same country in terms of its average temperatures, for a certain time period. In general, the approaches denoted by multirepresentation are based on extensions of the ER (Entity-Relationship) and UML (Unified Modelling Language) models in order to incorporate spatial and temporal features in the database modeling with different LoDs (Parent et al. 2009). Several data models, each one with specific concepts, have been proposed in the literature. In Parent et al. 2009 a survey about multirepresentation modeling is given in which three requirements are presented that should be verified in a multirepresentation approach. Firstly, a model should allow one to characterize the same object using different sets of attributes, or/and with different domain values. Secondly, a model should allow mapping one object to several objects or two different sets of objects. This is particularly useful when we change the spatial LoD, where objects may disappear and others may be grouped. Thirdly, a model should enable multiple representations of relationships. For instance, two regions might be modeled as spatially adjacent at a lower scale but at a more precise scale the regions are just near each other. According to Parent et al. 2009, MADS (Modeling Application Data with Spatiotemporal features) (Parent et al. 2006) is the only model which verifies the three requirements. It supports multiple spatiotemporal representations of a phenomenon mainly through perceptions. More particularly, we can assign perception stamps to any

element of the schema including objects, object attributes, and relationships. According to the perception stamp, we will have access to different spatial representations of objects or relationships, to different domain values of attributes, or even to different attributes.

Among the main drawbacks of multirepresentation is the fact that different LoDs, required by different applications, or the same application at different stages, can vary (Zhou et al. 2004). Bearing this in mind, the task of modeling a real-world phenomenon for which several spatial and/or temporal LoDs are needed can easily be very challenging. There are no pre-defined operations that take data from one spatial and/or temporal level to another. Everything is defined by the user at the instances level. Despite these drawbacks, having several pre-computed representations when dealing with numerous spatiotemporal events can be advantageous.

2.3.2 Multiresolution Approaches

Unlike the multirepresentation approaches, the multiresolution is essentially focused on the spatial component of the data. Plus, it derives the proper LoD on demand (Zhou et al. 2004). Data are stored at the highest level of resolution (or detail) and are dynamically generalized to lower LoDs, using known and pre-defined generalization operations. The generalization of spatial data is a non-trivial task and involves object simplification (e.g., at less precise resolutions, a building may be defined using less vertexes than originally), dimensionality reduction (e.g., a building can be represented by a polygon at a precise resolution, and by a point at a less precise resolution) and existence (e.g., eventually to represent that building is not relevant anymore). This sequence of operations was coined by Laurini 2014 as Generalization-reduction-disappearance process. More details about generalization operators can be found in (Weibel and Dutton 1999).

In an early work, Stell and Worboys 1998 define resolution or granularity (for the authors these are synonyms) as the level of discernibility between elements of a phenomenon that is being represented by the dataset.

Based on the resolution definition, Stell and Worboys 1998 define a stratified map space which consists of a set of maps representing the same spatial extent at different resolutions related to form a resolution lattice through general conversion operators (generalize and lift operators). Each map holds the same semantic and spatial granularity which corresponds to a database state. Maps are grouped by map spaces, i.e., sets of maps at the same resolution, describing the set of all possible databases states that are instances of some fixed schema. Through this work, the authors do not aim at a formalization of the complex process of cartographic generalization, but a framework as basis reasoning on generalized maps.

Brahim et al. 2015 propose a mathematical framework for the generalization of regions and ribbons (in vector space). The authors specify rules in order to specify when a ribbon turns into a line, an area becomes a point, and when the disappearance of a spatial object occurs. Furthermore, topological relations between ribbons are formalized as well

as between ribbons and regions. Then, the authors specify transformation rules that stipulate when topological relations between ribbons or between ribbons and regions need to be changed. On the contrary, in this work, a change in a topological relation between temporal or spatial features resulting from the generalization is not an imposition but rather a consequence of their individual generalization.

In (Zhou et al. 2004), a multiresolution approach to generalize polygonal data is proposed. The spatial generalization happens in a post-query process based on a scaless data structure. As to the time required to perform such operation, it is not clear. The authors make the following statement: "*We found that the overhead of simplify-while-retrieve approach based on the scaleless data structure is significant but not very large*". The generalization at run-time is important when such process depends on the data achieved at that moment which in turn may vary according to the user interaction like filtering over semantic attributes, spatial filters, and so on. The time required to perform the generalization process can be an issue. In an interactive application like VA approaches, the fast response time is crucial for the user and when dealing with numerous spatiotemporal events this is an open issue (Committee et al. 2013).

Moreover, to the best of our knowledge, the multiresolution approaches do not provide generalization operators that take into account the temporal component. Finally, they are more focused on the map visualization (and the corresponding spatial generalization operators) and less on the computation of data at different LoDs (Benz et al. 2004).

2.3.3 Granular Computing and Others Approaches

Bittner and Smith 2003 have developed a formal theory of granular partitions domain independent that uses "granules" (i.e., cells) to model abstract real-world entities at different granularities. The theory of granular partitions is bipartite: (*i*) the theory *A* characterizes the partitions as system of cells that are partially ordered by the subcell relation. As the partitions are cognitive devices that are directed towards reality, *ii*) the theory *B* defines their projective relation to the reality. Such theory brings the term object as any portion of the reality like an individual, a spatial region, a class of individuals. Then, an object can be recognized by some cell of a partition. A limitation of this theory is the lack of automatic methods to express a reality from one LoD to a coarser one.

A granular computing approach devised for spatiotemporal data was proposed by Camossi et al. 2006. The authors propose to represent spatiotemporal information (vector approach) in object-oriented database management systems (DBMSs) extending the ODMG standard. They define two new parametric data types. Spatial data types are defined through the $Spatial < G_s, \tau >$ data type, where G_s is a spatial granularity and τ being one of the ODMG types typically used to define conventional attributes like literal types (e.g., integer, float, etc.) or geometric types (like points, lines and polygons). Temporal or spatiotemporal data types are defined using the $Temporal < G_t, \gamma >$ data type

where G_t is a temporal granularity and γ can be any data type mentioned (including a spatial data type).

To $Spatial < G_s, \tau >$ and $Temporal < G_t, \gamma >$ data types, coarse and refinement functions can be assigned allowing to hold data at different granularities (i.e., several LoDs). Coarse functions convert data from a granularity G_α to a granularity G_β such that G_α is finer than G_β while refinement functions perform the opposite. We can have coarse or refinement functions applicable to spatial geometrical attributes or spatial quantitative and temporal attributes (Camossi et al. 2006). For example, coarse or refinement functions applied to spatial geometrical attributes can force some granules to modify their position and extent, be deleted, be split, and be merged. Some coarse functions that can be applied on numerical types are: min, max, average. Using this approach, the user specifies, for each class attribute, what conversion functions can be used (which are already defined by Camossi et al. 2006).

The $Spatial < G_s, \tau >$ data type indexes information of the type τ to spatial granules. Furthermore, the $Temporal < G_t, Spatial < G_s, \tau >>$ data type indexes the information of the type τ already indexed by spatial granules to temporal ones. Note that, when we define a temporal data type, the temporal granules are specifying the valid time of the information indexed to them. Another important aspect of this approach is that the indexed information will not be granules of some granularity but values of some type τ (belonging to some domain). As a result, in some scenarios, we cannot relate information at different LoDs. Consider the following class attributes: (*i*) $Spatial < G_{Countries}, int >$ storing information about the exact population number in each country; (*ii*) $Spatial < G_{Countries}, String >$ also storing information about the population number but with less precision so that the possible values are: (*i*) less than one million (*ii*) one million or more and less than fifteen millions; (*iii*) fifteen or more millions. Although both variables refer to the same information, we cannot relate them by stating that the former is finer than the latter. This kind of reasoning is also important to relate spatiotemporal data at different LoDs.

Bravo and Rodríguez 2014 propose a multi-granular database model and a query language in order to query data using different granularities. This was done bearing in mind that events may be stored at different LoDs with respect to time (i.e., day, month, season, year) and location (i.e., city, country, zone), and despite differences in data granularity, the objective was to retrieve data at a specific granularity. Relying on the concept of domain schema, the multi-granular database model and a query language address heterogeneity of granularities in data. This way, data can be collected and stored at different granularities, and at query time, the data are derived at a particular granularity, keeping the results consistent. To move data from finer granularities to coarser ones, (Bravo and Rodríguez 2014) follow the same generalization process regardless of wether this generalization of data is described by spatial attributes or temporal attributes, for instance. Furthermore, in their model, each attribute is described by a granule. This will rise an issue that will be discussed into detail in Section 4.1.

More recently, a compressed hierarchical data structure was proposed in order to hold spatiotemporal events at multiple LoDs Lins et al. 2013. They focus on providing real-time exploratory visualization for huge amounts of spatiotemporal events. The research of Lins et al. 2013 is aligned with the work here proposed as they have available spatiotemporal events at different LoDs. However, each LoD is computed purely based on aggregation operators (e.g., count, max, min) and the generalization-reduction-disappearance process is not considered.

In short, these approaches are modeling spatiotemporal data at different LoDs by indexing and aggregating spatiotemporal data at different LoDs. We aim to go a step further and represent spatiotemporal phenomena logged as spatiotemporal events at different LoDs, including the generalization-reduction-disappearance process so that each representation of phenomenon at a particular LoD is computed following a bottom-up automated approach.

2.4 Manifold LoDs Approaches

The scale (or LoD) of analysis can greatly affect results (e.g., Modifiable Areal Unit Problem - MAUP). This issue has been acknowledged a long time ago (Openshaw and Openshaw 1984). However, with spatiotemporal events in mind, analytical approaches have been mainly developed to support analyses based on a single LoD. Thus, the MAUP becomes a problem, once unsuitable LoDs can hide patterns and conceal the true underlying nature of a dataset.

The LoD of analysis can affect results and this can be seen as an opportunity to develop approaches that work at different LoDs. VA approaches working across LoDs are still in its infancy despite the fact that they have been gaining more attention in recent years. We found some ad-hoc approaches working on multiple LoDs concerning spatiotemporal data but bearing a specific analytical goal.

Camossi et al. 2008 propose a spatiotemporal clustering technique applicable to different temporal and spatial LoDs in order to improve a clustering algorithm efficiency. The appropriate temporal and spatial LoD depends on a trade-off between the mining efficiency and the maximum detail desired, which is an input parameter. The choice of the temporal and spatial LoD is done iteratively through the LoDs available until the best trade-off is found.

Malizia and Mack 2012 enhanced the Jacquez k nearest neighbor test in order to identify the spatial and temporal LoDs at which spatiotemporal interaction takes place.

ArcGIS is currently one of the most widely used commercial GIS software for working with maps and geographic information. It has an *Incremental Spatial Autocorrelation* tool which applies the Global Moran's I for a series of distances (i.e., different LoDs). Significant peak values suggest the spatial LoDs where the clustering is most pronounced, and therefore, the spatial LoDs that are more appropriate for investigating hotspots.

Watson 2015 developed a visualization method which displays n events across multiple temporal LoDs within a single image. Time maps are built as follows. Imagine each event as a dot along a time axis. Then, the time differences between consecutive events (in time) are computed. The time map is a scatter plot, where the (x, y) coordinates are specified by each neighboring pair of time delays. Every point drawn on a time map corresponds to an event within the dataset. In Figure 2.11α, two sequences of events are displayed and their corresponding time maps. In sequence A, the events are evenly spaced. In the corresponding time map, all of the points are exactly on top of each other. Sequence B is similar to sequence A but the timing of one event (the red) was changed. To simplify the interpretation of time maps, the author provided a heuristic diagram that is divided into four quadrants (see 2.11β).

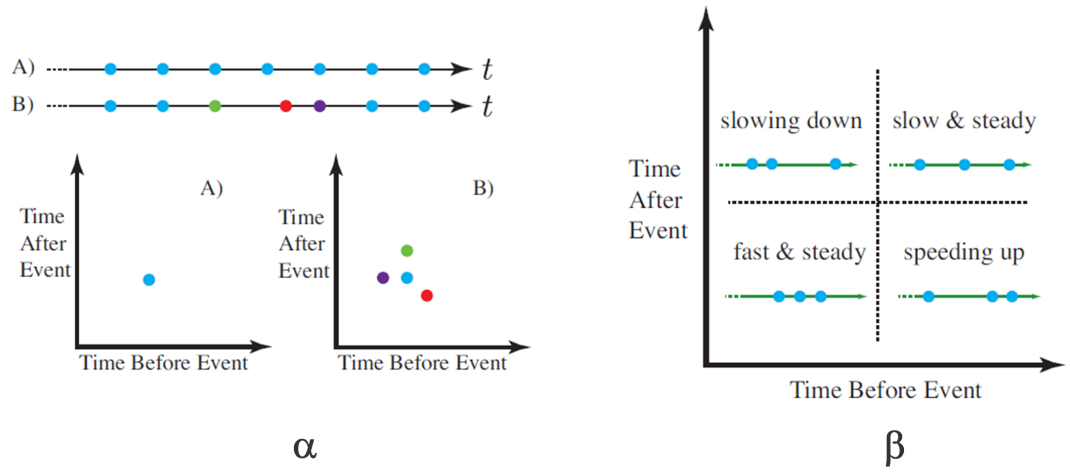


Figure 2.11: Illustration of the Time Maps visualization and their heuristic interpretation Watson 2015.

Roughly speaking, events in the lower-left and upper-right quadrants are regularly spaced in time between the events directly before and after them. The events in the lower left occur in rapid succession while at the upper-right they happen at a slower rate. Events in the upper-left quickly follow their preceding event, and a longer time elapses until the next event. The lower-right quadrant is similar to "speeding up" since a long delay is followed by an accelerated pair of events.

The authors use their approach on the 3.200 most recent tweets written by Barack-Obama (likely president staff). The tweets occurred from October 2013 to April 2015. A heated time map was produced that can be seen in Figure 2.12. Two main patterns are easily recognized at different temporal LoDs. During major events like the 2015 State of the Union Address, a tweet is written every few minutes. On other days, the tweet rate is about one per hour. This kind of insights in several LoDs from events are the object of this work. However, this approach is just dealing with the temporal dimension of events.

Sips et al. 2012 propose a Visual Analytics approach called *Pinus*, aiming at the detection of patterns at multiple temporal LoDs in numerical time series, specifically

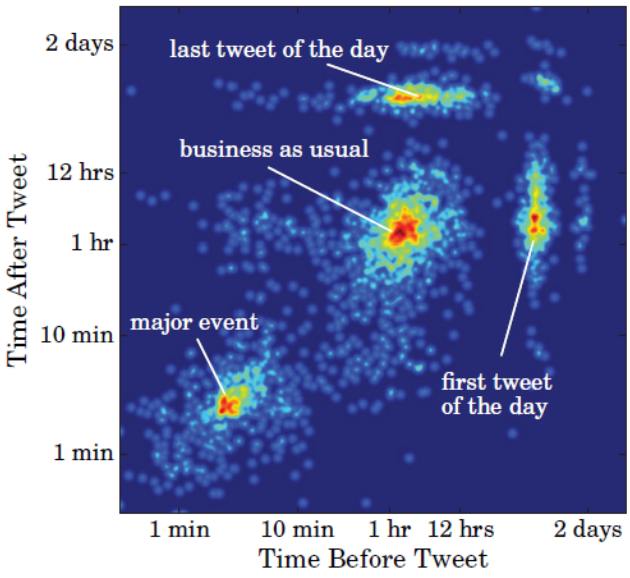


Figure 2.12: A heated time map for tweets written by @BarackObama (Watson 2015).

from environmental sciences. To accomplish that, statistical measures are computed for all possible time LoDs (i.e., scales) and starting positions, namely, mean, variance, and discrete entropy were implemented.

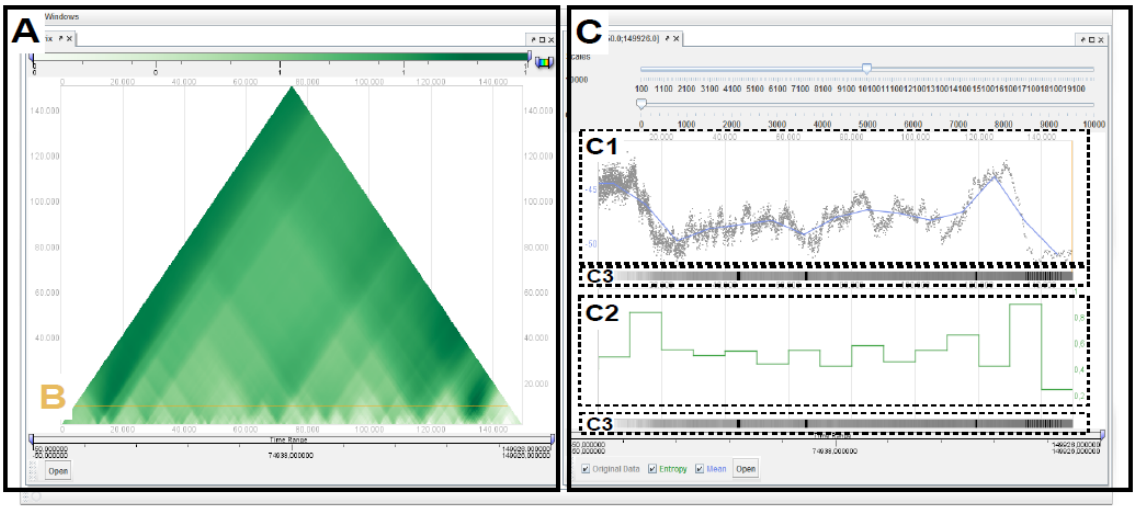


Figure 2.13: An overview of the Pinus View prototype proposed by (Sips et al. 2012).

An overview of the prototype developed is given in Figure 2.13. The *Pinus view* is the starting point for detecting patterns (see Figure 2.13A). This particular case is showing the variation of the entropy for many time scales and time steps so that whites map zero entropy while dark greens map maximum entropy. Users can select a temporal LoD directly in the Pinus view as illustrated as *B* in Figure 2.13A. Panel C (right) shows the result of the query, i.e., Panel C1 shows the original data points of the time series. Panel C2 shows the entropy values at 10k year temporal LoD. Notice that, the data displayed in

Figure 2.13 was used to gain insights about the glacial climate record data derived from an ice core from Dronning Maud Land, Antarctica that were presented in Section 1.1.

This approach makes no assumption about the temporal LoD and the temporal patterns. It combines statistic measures and the pattern recognition abilities of the user to support effective detection of temporal patterns at different temporal LoDs. We aim to bring this mindset for the analysis of spatiotemporal events at several spatiotemporal LoDs.

Goodwin et al. 2016 propose a framework for analyzing multiple variables across spatial LoDs and geographical locations. Based on it, they developed a suite of novel interactive visualization methods to identify interdependencies in multivariate data coupled with a series of correlation matrix views. An overview of the interface is given in Figure 2.14.

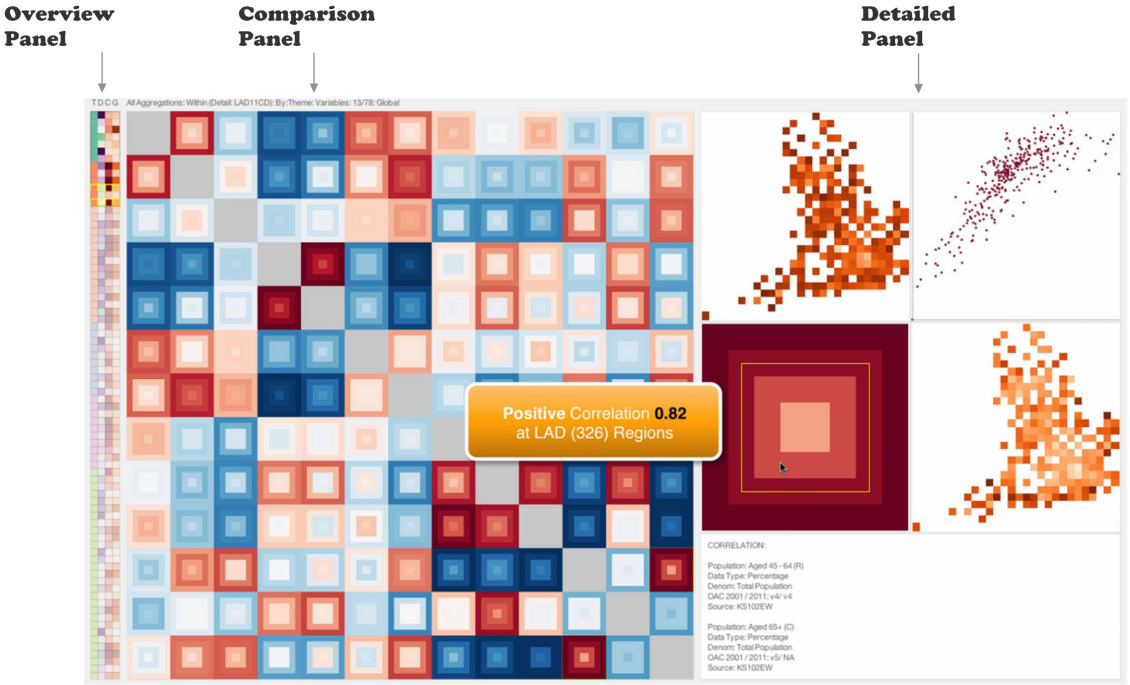


Figure 2.14: An overview of the Goodwin et al. 2016 approach.

The overview panel allows all variables to be ranked by four global measures: theme (variable category), skewness (as an indication of distribution), variance of correlation (as an indication of how correlation varies across all variables) and Moran's I (to establish geographical dependencies).

The comparison panel is an adjustable correlation matrix. Figure 2.14 is filled with the scale mosaic matrix proposed by the authors. Rows and columns represent variables ordered according to the overview panel. Each cell is multicolored (based on a partitioned square) in order to represent the correlation of a pair of variables at different spatial LoDs. As the partition is closer to the center, the finer the spatial LoD is (e.g, administrative levels – state, county, district, census level). The detail panel shows details concerning

the spatial LoD chosen in the bottom-left visualization. It might contain maps showing the geographical distribution of the individual variables or pairwise local correlation, as well as a scatterplot presenting the local correlation.

In the example shown in Figure 2.14, the variables at study are the percentage of population aged 45-64 and the percentage of population aged 65+. These variables reveal that positive correlation decreases as long as we consider lower spatial LoDs.

This approach does not focus on a particular phenomenon and was devised to look for correlations on multiple variables in multiple spatial LoDs and geographic regions.

Robinson et al. 2016 developed a visual analytics approach, called STempo, to support the discovery of patterns found in spatiotemporal events. STempo was designed to detect and analyze significant co-occurrences of real-world events. The dataset used was carefully prepared and extracted from internet news feeds. Each event corresponds to a single news event that contains information about its type (e.g., Political, Diplomatic), the latitude and longitude coordinates, the date, among other information. The events collected take place in Syria.

STempo includes several coordinated views as can be seen in Figure 2.15. STempo leverages T-pattern (Magnusson 2000), a method for identifying sequences of significantly co-occurring events. The sequences revealed by T-pattern analysis (i.e., the co-occurring

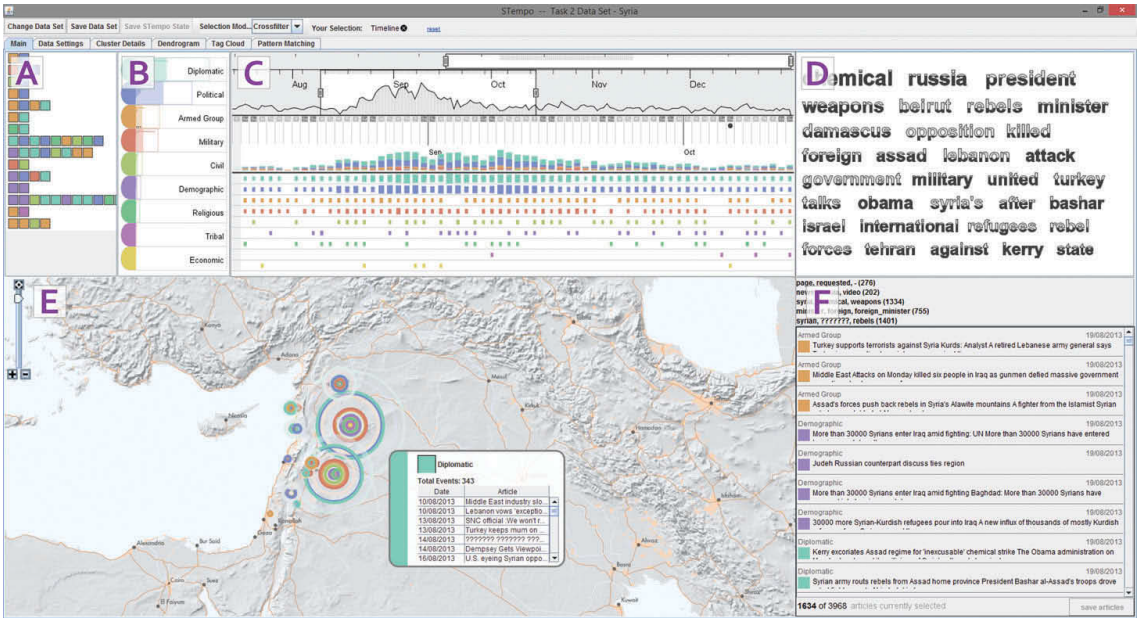


Figure 2.15: An overview of the STempo (Robinson et al. 2016).

event types next to one another over time) are shown in Figure 2.15a ordered by their statistical significance. They are shown in a simplified way with colored blocks to represent each of the nine high-level event categories that are displayed in Figure 2.15b which in turn are ordered by the frequency of occurrence for each category. STempo also has a timeline tool (Figure 2.15c) that displays the number of events and allows users to focus on a particular time interval. A tag cloud (Figure 2.15d) shows the key words used in the

titles of articles from the dataset at study. An interactive map shows events located on a custom-designed terrain basemap using colored concentric rings to represent associated higher-level event categories (Figure 2.15e).

This approach was developed to make T-pattern analysis over spatiotemporal events regarding news taking place in Syria. This is done for all the events and, at this moment, temporal or spatial filters cannot be applied as a way of changing the input data of the T-pattern analysis. Furthermore, this approach is making a separate analysis of the temporal and spatial dimension of events as the input for the T-pattern algorithm corresponds to records containing the timestamp and a set of event types that occurred in it. Finally, this approach looks for temporal patterns and not for spatiotemporal patterns, because the sequences identified are not assigned to specific geographic regions, for instance. Nevertheless, this approach computes temporal patterns in multiple temporal LoDs because each sequence identified is anchored to a specific temporal LoD.

The visual analytics approaches discussed so far explore time following a linear model. However, periodicity is underlying in all societies. Examples of periodicity can be seasonal changes in the weather, Ramadan, our monotonous daily tasks, among others. However, different calendars (e.g. Islamic, Gregorian), different cultural backgrounds, and other variables encumber the analytical ability to uncover and understand human activity at a given time within a specified region. In order to address the periodicity and the calendar heterogeneity, Swedberg and Peuquet 2016 propose a visual analytics web application developed to help users in the detection and analysis of calendar related periodicity in spatiotemporal event data sets via exploratory user interaction.

An overview of the PerSE's interface is given in Figure 2.16. The main views of the interface are the map (Figure 2.16B), the attribute (Figure 2.16C), the time-wheel (inspired in the CircleView) (Figure 2.16D), the timeline (Figure 2.16E) and the table (Figure 2.16F). These views are coordinated - the different views are displaying the same data but from different perspectives. Furthermore, an interaction in a particular view like choosing a period of time has an effect in the others views. Finally, the metric displayed is the events frequency.

The map can be divided into a maximum of six geographic regions, and for each one, there is a time-wheel. In Figure 2.16, the time-wheels use the Gregorian Calendar but the calendar can be changed for the Islamic calendar, for instance.

This work allows for the analysis at multiple spatial LoDs and temporal LoDs despite the fact that the number of the spatial LoDs that we can analyze, simultaneously, are limited to two (raw data and aggregated by the user-defined geographic regions). The authors illustrate their approach using a subset of the Nigerian dataset (number of events=4,854) taken from the ACLED (Armed Conflict Location and Event Data Project). Examples of patterns detected by Swedberg and Peuquet 2016 are:

- *A day-of-week pattern only evident within northern, central, and western Nigeria. The pattern suggests that Sunday through Tuesday have a higher frequency of violence.*

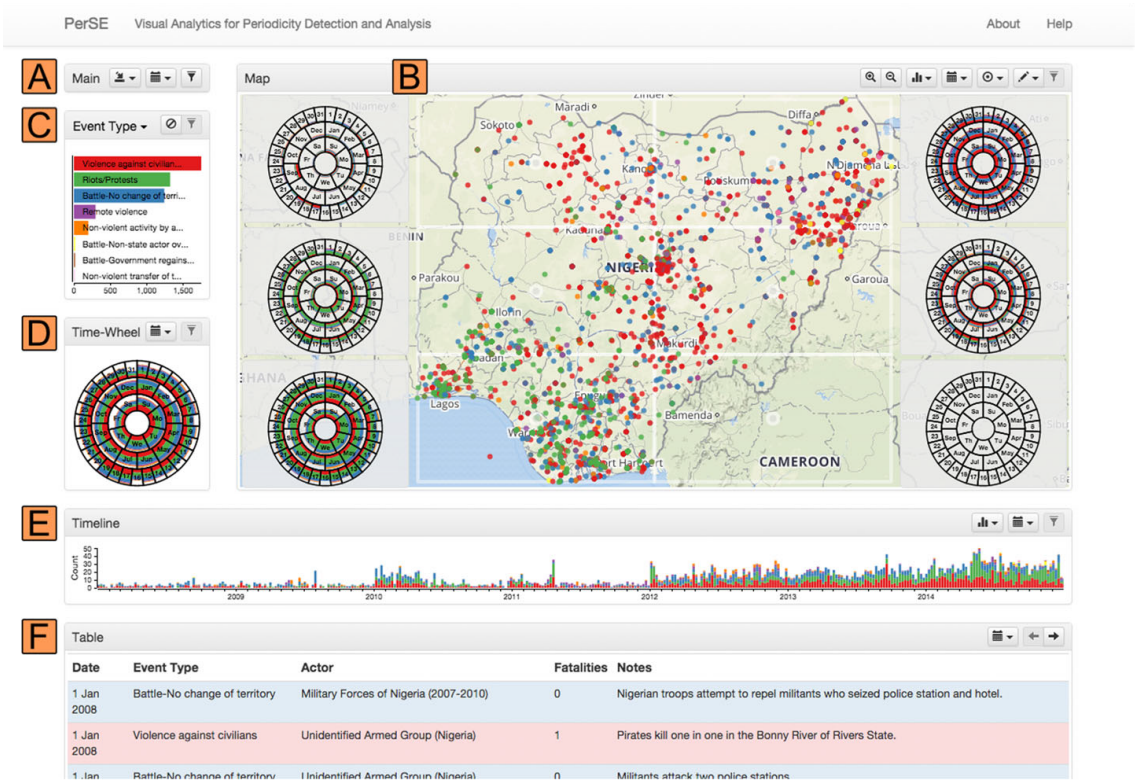


Figure 2.16: An overview of the PerSE prototype proposed by (Swedberg and Peuquet 2016).

- *A month-of-year pattern in the Gregorian calendar within the northeastern Nigeria. The pattern suggests that January, February, and March contain less frequent violent events.*

- *A month-of-year pattern in the Islamic calendar within northeastern Nigeria. The pattern suggests that Boko Haram is more active around the months of Ramadan, Shawwal, and Muharram.*

Although the mentioned patterns are interesting, they are obtained by working with space and time separately using only the descriptive statistic COUNT. As pointed out in the beginning of this PhD Thesis, much information might become visible when one works with the spatial and temporal dimensions together.

To the best of our knowledge, there are no approaches that work across several spatial and temporal LoDs, working with space and time together, and therefore, looking for spatiotemporal patterns at different spatiotemporal LoDs. Furthermore, the VA approaches discussed do not have any theoretical foundation that anchors the analysis across LoDs. The approaches rely on clever visual designs that show data at different LoDs. However, from our perspective, a theoretical foundation that anchors the analysis across LoDs can be important for having phenomena representations for different LoDs, and then, use better suited visualization methods to display them.

Approaches working across several LoDs are needed and, as shown, they are starting to be developed. This work seeks for an approach that follows the VA Mantra without focusing on a particular analytical task/pattern, which can be applicable in the context of spatiotemporal events.

Chapter 3

Theory of Granularities

Humans are constantly using granularities in unconscious ways, in order to make statements about phenomena. Those granularities have an underlying domain of reference. In most cases, granularities are just a way to create a domain of discourse simpler than their domains of reference. This can be observed when we use several levels of administrative divisions to make it easier to refer a to particular country area; it can be observed when we refer to time as days, or months; it can be perceived when we assign the age of a person always rounded to units; and these are just a few examples. Here, we denote a domain of reference of a granularity as $D = (DS, RS)$ where the domain set DS corresponds to a set of elements and RS is a set of relations defined over DS. A domain set can be discrete, dense, continuous or n-dimensional. A granularity is formally defined as follows.

Definition 3.1 (Granularity)**.** Let $\mathcal{IS}$ be an index set; $D = (DS, RS)$ be a domain; 2^{DS} the power set of the DS; and GS be a subset of 2^{DS} apart from the empty set, $GS \subseteq 2^{DS} \setminus \{\emptyset\}$ such that all elements of GS are disjoint from each other. A granularity G is a bijective mapping from GS to the index set $\mathcal{IS}$:

$$G : GS \rightarrow \mathcal{IS} \tag{3.1}$$

A granularity G defines a division of a domain in a set of granules. A granule g_{ind} corresponds to a pair (g, ind) where $g \in GS$ and $ind \in \mathcal{IS}$. The extent of the granule g_{ind} is denoted by $E(g_{ind})$ which is g; the index value of the granule g_{ind} is denoted by $I(g_{ind})$ which corresponds to ind. The set of extents of granules is denoted by $GrS(G)$. The union of elements belonging to $GrS(G)$ defines the extent of a granularity $Ext(G)$.

Let's consider the *white area* within the ellipse displayed in Figure 3.1 as our domain to be granulated. In this domain, an extent of a granule is represented by a pink area and the corresponding index value is a letter over the pink area. This way, g_1 is an example of a granule. The granules $g_1, g_2, \ldots, g_7$ define a granularity over the *white area*.

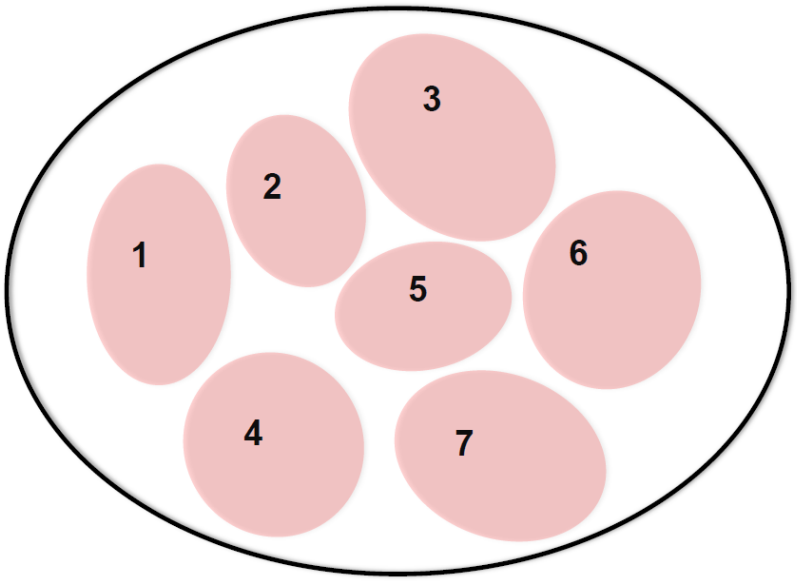


Figure 3.1: Illustration of the granularity concept.

Unlike the majority of the proposals that can be found in the literature, we propose a mapping from the granules to an index set rather the reverse (see Section 2.2). Using a mapping from the index set to granules can lead to too many values from the index set mapped to the empty set. Using this granularity definition, we only need to define the set of extents of granules and their corresponding mapping to the index set.

There are constraints concerning the mapping between the set of granules GS and the index set $\mathcal{IS}$. Through a bijective mapping, we are imposing the following constraint: every element of GS is mapped to exactly one element of $\mathcal{IS}$. Consequently, different granules cannot share an index value, and one granule cannot be associated to more than one index value.

A granularity defines a set of granules that can be used to refer to a particular domain with a certain level of abstraction. Granules might represent familiar concepts for us (as humans) or not, i.e., they are just a portion of the granulated domain. Through the granularity definition proposed, it is possible to define a granularity over any domain including the ones proposed in the literature (see Section 2.2). This is important to model spatiotemporal events at different LoDs later on as they are described through different domains of reference like space domain, time domain, among others.

Let's consider events about forest fires in Portugal[1] such that for each incidence the location, time and its cause is described. These attributes have different domains of reference and will be used to illustrate the concept of granularity.

A possible domain of reference for time is the domain of real numbers with total order $D_1 = (\mathbb{R}, <)$. The data about forest fires are provided based on a granularity $Minutes$ where each granule represents a minute. The exact minute at which the fire starts can

[1]Data provider: `http://www.icnf.pt/portal/florestas/dfci/inc/estat-sgif`

be irrelevant. One may want to analyze the hour at which the forest fires have happened. Thus, the granularity $Hours$ over the domain D_1 where each granule represents an hour can be defined. However, the hour may be too detailed to get an insight about the time of day that most forest fires begun. Thus, we can consider the granularity $DaysSubUnits$ with granules representing the several periods of each day: night, morning, heat's peak, and afternoon. Additionally, it can be interesting to analyze in what days the forest fires happened. The granularity $Days$ should be defined where each granule represents a day, illustrated in Figure 3.2. Finally, the appropriate granularity depends on the phenomenon and the analytical goal. These examples of granularities correspond to temporal granularities proposed by Bettini et al. 2000.

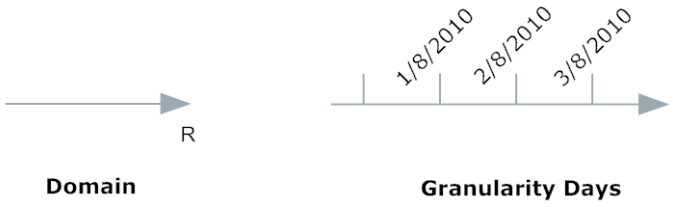


Figure 3.2: Example of a granularity defined over D_1.

A possible domain of reference to describe the Earth's surface can be the two-dimensional coordinates with fourteen decimal cases $D_2 = (\mathbb{R}^2, RS)$, with a set of relations RS irrelevant for the following examples. The data about forest fires are provided based on the domain D_2. An analyst may want to identify what parishes in Portugal present a larger burnt area. In this case, it is necessary to define the granularity $Parishes$ where each granule refers to a Portuguese parish. Moreover, an analyst might be interested to know in what of Portugal's districts more forest fires occur. For this scenario, let's assume that we need the granularity $Districts$ where each granule corresponds to a district of Portugal (see Figure 3.3). This granularity is defined over D_2. Also, the identification of protected areas in which there is more burnt area can be desirable. In this case, we need the granularity $ProtectedAreas$ where each granule refers to a protected area in Portugal. Again, the appropriate granularity depends on the phenomena and the analytical goal. The granularities $Parishes$, $Counties$ and $ProtectedAreas$ correspond to the spatial granularities proposed by Camossi et al. 2006.

Through the granularity definition proposed, we can also define granularities over domains unrelated with time domains or space domains. One of the attributes describing a forest fire incident is the forest fire's cause. The domain of reference of this attribute is discrete and contains a long list of possible causes, namely stubble burnings, electrical power, mental illness, and so on. The LoD underlying this domain may be too detailed if an analyst is just interested in discerning the places where there are more forest fires caused by accident from the places where the occurrence of forest fires intentionally caused by humans is usual. In this case, a granularity $Causes$ composed by three granules should be defined. One that encompasses unintentionally and indirectly human causes

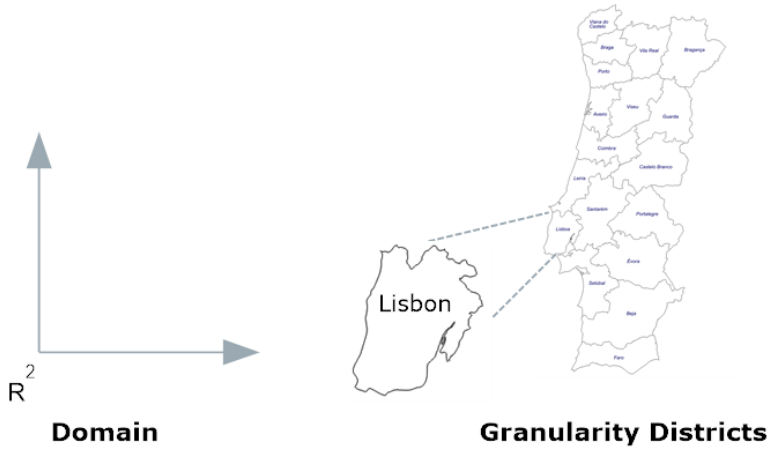


Figure 3.3: Example of a granularity defined over D_2.

($causes_{use\ of\ fire}$), another that embraces accidental causes ($causes_{accident}$) and another one that covers intentional human causes ($causes_{voluntary}$). This granularity is illustrated in Figure 3.4 so that on the left side the domain of the forest fire's cause the attribute is shown whereas on the right side the granularity *Cause* is displayed. This kind of granularity is only supported by Keet's theory (Keet 2008).

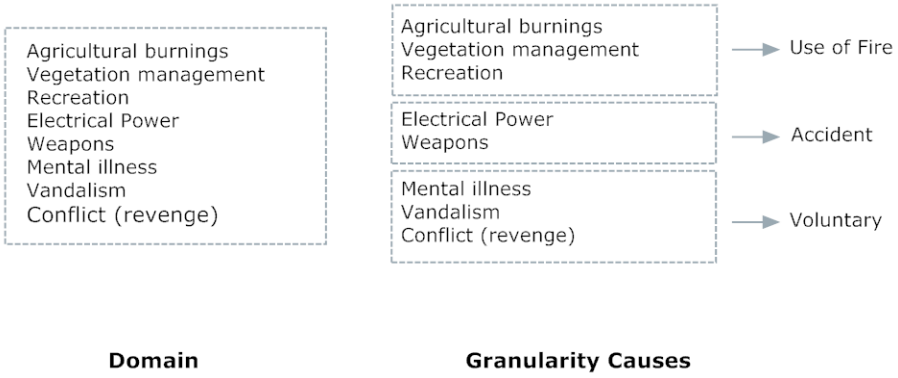


Figure 3.4: Example of a granularity defined over the *cause* attribute provided by the data provider.

The Normalized Difference Vegetation Index (NDVI) is produced by the Moderate Resolution Imaging Spectroradiometer (MODIS) aboard NASA's Terra satellite[2]. This index is a measurement about vegetation on Earth. The raster data are provided every 16 days at 250 meter spatial resolution. To each cell the NDVI value represents the entire period (16 days) and the corresponding area. Index values ranging from 0.4 to 0.9 mean lands covered by vegetation while lower values (0 to 0.4) mean lands where there is little or no vegetation.

The data is provided based on a temporal granularity where each granule refers to a 16 days time period. This granularity may be too detailed if a user wants to monitor

[2] https://neo.sci.gsfc.nasa.gov/view.php?datasetId=MOD13A2_M_NDVI

and investigate shifts in plant growth patterns that occur in response to climate changes. Thus, the granularity $Year$ where each granule refers to a period of a year can be sufficient to analyze such changes.

On the other hand, the data is provided based on a "raster" spatial granularity, i.e., all granules are areas of 250 meters by 250 meters. This spatial granularity can be too detailed if a user wants to analyse shifts in plant growth patterns for the entire Earth surface. For such a scenario, the granularity $Raster(10km^2)$ where each granule refers to an area of 10 km^2 may be sufficient. These granularities are similar to the ones proposed by Pozzani and Zimányi 2012.

The granules of a granularity can be related to each other through relationships. We introduce the possibility to annotate a granularity in order to define relations between granules of a granularity. An annotation over a granularity G corresponds to a binary relation defined on the set of granules.

A granularity annotation can be useful in any granularity defined over any domain. Recall the granularity $Days$, where each granule refers to a period of a day. This granularity can be annotated with the relationship next working day. Now, consider the granularity $Countries$ where each granule refers to a particular country. This granularity can be annotated in order to specify what countries hold privileged trade relations, alliances or conflicts between them; relations of exporter/importer of oil, natural gas, gold, and other materials.

3.1 Reasoning over Granules

3.1.1 Relations between Granules

There are known relationships in the domains that we would like to preserve or transpose to the granularities. For example, Bettini et al. 2000 are interested in temporal granularities where the granules are totally ordered, which is related with the total order underlying the time domain. To guarantee that, Bettini et al. 2000 introduce a set of constraints in the temporal granularity definition as presented in Section 2.2.

Likewise, and regardless of the domain, we may be interested to bring relations defined in the original domain to the granules. We propose four ways to transpose a relation, defined in the domain of a granularity, for two granules of such granularity. Therefore, we introduce four relations that can be defined between granules of a granularity. The relationships proposed are: (*i*) complete; (*ii*) partial; (*iii*) weak; (*iv*) and, existential. These relationships are induced from the relations held by the elements of the domain of a granularity.

Given a granularity G defined over a domain $D = (DS, RS)$, a relation R defined over DS such that $R \in RS$, and g_i and g_j denotes two granules belonging to G. The formal definitions of the relationships are given.

Definition 3.2 (Complete Relationship)**.** A complete relationship $g_i \; R^C \; g_j$ is defined as follows.

$$g_i \; R^C \; g_j \Leftrightarrow \forall x_i \in E(g_i), \forall x_j \in E(g_j) : x_i \; R \; x_j \tag{3.2}$$

If two granules g_i and g_j are completely related then all elements of g_i must be related with all elements of g_j through the relation R.

Definition 3.3 (Partial Relationship)**.** A partial relationship $g_i \; R^P \; g_j$ is defined as follows.

$$g_i \; R^P \; g_j \Leftrightarrow \exists x_i \in E(g_i), \forall x_j \in E(g_j) : x_i R x_j \wedge \exists x_j \in E(g_j), \forall x_i \in E(g_i) : x_i R x_j \tag{3.3}$$

In case of two granules g_i and g_j are partially related then there is at least one element in g_i related with all elements of g_j through the relation R and similarly, there is at least one element in g_j where all elements of g_i are related with g_j through the relation R.

Definition 3.4 (Weak Relationship)**.** A weak relationship $g_i \; R^W \; g_j$ is defined as follows.

$$g_i \; R^W \; g_j \Leftrightarrow \exists x_i \in E(g_i), \forall x_j \in E(g_j) : x_i R x_j \vee \exists x_j \in E(g_j), \forall x_i \in E(g_i) : x_i R x_j \tag{3.4}$$

When two granules g_i and g_j are weakly related then there is at least one element in g_i related with all the elements of g_j through the relation R or, there is at least one element in g_j where all elements of g_i are related with g_j through the relation R.

Definition 3.5 (Existential Relationship)**.** An existential relationship $g_i \; R^E \; g_j$ is defined as follows.

$$g_i \; R^E \; g_j \Leftrightarrow \exists x_i \in E(g_i), \exists x_j \in E(g_j) : x_i \; R \; x_j \tag{3.5}$$

Finally, for two granules g_i and g_j to be existentially related, at least one element of each granule is related via the relation R.

(a) $g_a \; R^C \; g_b$ (b) $g_a \; R^P \; g_b$ (c) $g_a \; R^W \; g_b$ (d) $g_a \; R^E \; g_b$

Figure 3.5: Illustration of the induced relations

In order to illustrate the four relationships proposed, consider a granularity S defined over the domain $D_3 = (\mathbb{R}^2, north)$ such that a coordinate (x_i, y_i) is at north of a coordinate (x_j, y_j) if and only if $y_i > y_j$. In Figure 3.5, there are four scenarios of two granules g_a and g_b belonging to S available. In Figure 3.5a, g_a is completely north of the granule g_b

(g_a $north^C$ g_b) since all elements of g_a are north of all elements of g_b. Looking at Figure 3.5b, g_a is partially north of the granule g_b (g_a $north^P$ g_b). In this case, there are some elements of g_a north of all elements of g_b and, there are some elements of g_b for which all elements of g_a are north. Regarding the Figure 3.5c, g_a is weakly north of the granule g_b (g_a $north^W$ g_b) because there are just some elements of g_a north of all elements of g_b. Finally, considering the Figure 3.5d, g_a is existentially north of the granule g_b (g_a $north^E$ g_b) since there are some elements of g_a north of some elements of g_b.

The induced relations are successive relaxations, i.e., $g_i\ R^C\ g_j \Rightarrow g_i\ R^P\ g_j \Rightarrow g_i\ R^W\ g_j \Rightarrow g_i\ R^E\ g_j$. Therefore, given a granularity G, the induced relations transpose how strong a relation R, defined over the DS, is verified between two granules.

Furthermore, it is important to know what properties of relations defined over the DS are preserved in the induced relations. For that, we consider five properties that a relation R can hold: (*i*) symmetric; (*ii*) transitive; (*iii*) reflexive; (*iv*) antisymmetric; (*v*) antireflexive.

Table 3.1: The induced properties of relations based on the properties of relations in the domain.

	$g_i\ R^C\ g_j$	$g_i\ R^P\ g_j$	$g_i\ R^W\ g_j$	$g_i\ R^E\ g_j$
Symmetric	√	√	√	√
Transitive	√	√	*inconclusive*	*inconclusive*
Reflexive	*inconclusive*	*inconclusive*	*inconclusive*	√
Antisymmetric	√	√	*inconclusive*	*inconclusive*
Antireflexive	√	√	*inconclusive*	*inconclusive*

It can be proved that if the relation R is symmetric then any induced relation is also symmetric. Furthermore, if the relation R is transitive then we only can state that the complete and partial relations are also transitive. For the other relations, nothing can be stated. Regarding the property reflexivity, only the existential relation is in any case also reflexive. Finally, if the relation R is antisymmetric or antireflexive then the complete and partial relations are also antisymmetric or antireflexive, respectively. The summary of these results is displayed on Table 3.1.

To achieve the results displayed in Table 3.1, formal demonstrations were conducted in the natural deduction system. Here, we discuss into more detail the transitive property and all the others are available in Appendix A. A relation R of a domain D is transitive whenever verifies the following property: $\forall x \forall y \forall z((x\ R\ y \wedge y\ R\ z) \rightarrow x\ R\ z)$. If an element x is related to an element y through the relation R and y is related to an element z via relation R then x and z are also related through the relation R. Consider, the previous defined domain D_3. The north relation is transitive because if a coordinate c_1 is north of a coordinate c_2 and the coordinate c_2 is north of a coordinate c_3 then the coordinate c_1 is north of the coordinate c_3.

According to the formal proof provided in Fitch-style calculus, available in Figure 3.6,

given a domain D and a relation R, if the relation R is transitive on the domain D then we can state that the complete relationship R^C, between granules belonging to a granularity defined over D, is also transitive. In addition to g_i and g_j, consider also a granule g_k belonging to S. We take as premises a transitive relation R, $g_i\ R^C\ g_k$ and $g_k\ R^C\ g_j$ and we want to proof $g_i\ R^C\ g_j$. By taking arbitrarily three elements (line 4^{th}) and using universal instantiation, we can infer that an element of g_i is related with an element of g_k: $a\ R\ b$ (line 5^{th}); and an element of g_k is related with an element of g_j : $b\ R\ c$ (line 6^{th}). Since the relation R is transitive we can infer $a\ R\ c$ (line 9^{th}). Through universal introduction, we can conclude that any element of a granule g_i is related to g_j via the relation R (line 10^{th}).

1. $\forall x\ \forall y\ \forall z\ ((R(x,y) \land R(y,z)) \rightarrow R(x,z))$	
2. $\forall i\ \forall k\ R(i,k)$	
3. $\forall k\ \forall j\ R(k,j)$	
4. $\boxed{a, b, c}$	
5. $R(a,b)$	$\forall$ **Elim:** 2
6. $R(b,c)$	$\forall$ **Elim:** 3
7. $R(a,b) \land R(b,c)$	$\land$ **Intro:** 5, 6
8. $(R(a,b) \land R(b,c)) \rightarrow R(a,c)$	$\forall$ **Elim:** 1
9. $R(a,c)$	$\rightarrow$ **Elim:** 7, 8
10. $\forall i\ \forall j\ R(i,j)$	$\forall$ **Intro:** 4–9

Figure 3.6: A transitive relation induces transitive complete relationships.

A particular example can be observed in Figure 3.7a. The granules g_u, g_v, g_w are granules of a granularity defined over $D_3 = (\mathbb{R}^2, north)$. The granule g_u is completely north of the granule g_v ($g_u\ north^C\ g_v$) and the granule g_v is completely north of the granule g_w ($g_v\ north^C\ g_w$) then the granule g_u is completely north of the granule g_w ($g_u\ north^C\ g_w$).

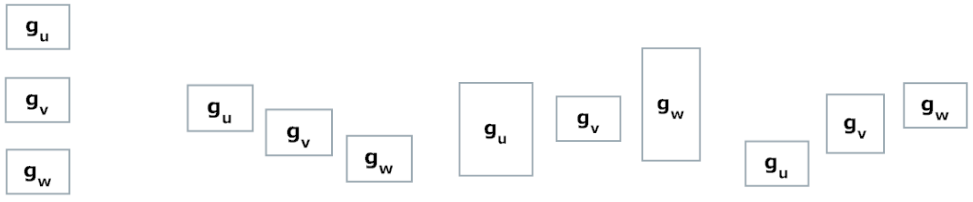


Figure 3.7: Four scenarios for granules belonging to S.

A similar statement can be made regarding the partial relationship R^P. In short, a transitive relation R induces a transitive partial relationship R^P as shown in the proof available in Figure 3.8. We take as premises a transitive relation R, $g_i\ R^P\ g_k$ and $g_k\ R^P\ g_j$ and we want to proof $g_i\ R^P\ g_j$. Based on the 2^{nd} and 3^{rd} premises, we can infer that there is at least one element in g_i related to all elements of g_k through the relation R (line

4^{th}); and we can infer that there is at least one element in g_k related with all elements of g_j via the relation R (line 5^{th}). It can be concluded that there is at least one element in g_i related with all elements of g_j through the relation R (from 6^{th} line to 17^{th}). This corresponds to the left-hand side of the conjunction that defines the partial relationship $g_i \; R^P \; g_j$. Then, a similar reasoning was performed looking at the right-hand side of the conjunction (from 18^{th} line to 31^{th}). Through the conjunction introduction of the two intermediate conclusions (line 32^{th}) we got $g_i \; R^P \; g_j$.

1. $\forall x \; \forall y \; \forall z \; ((R(x,y) \wedge R(y,z)) \rightarrow R(x,z))$
2. $\exists i \forall k \; R(i,k) \wedge \exists k \forall i \; R(i,k)$
3. $\exists k \forall j \; R(k,j) \wedge \exists j \forall k \; R(k,j)$

4. $\exists i \forall v \; R(i,k)$ — $\wedge$ **Elim:** 2
5. $\exists k \forall j \; R(k,j)$ — $\wedge$ **Elim:** 3
   6. [a] $\forall k \; R(a,k)$
      7. [b] $\forall j \; R(b,j)$
      8. $R(a,b)$ — $\forall$ **Elim:** 6
         9. [c]
         10. $R(b,c)$ — $\forall$ **Elim:** 7
         11. $R(a,b) \wedge R(b,c)$ — $\wedge$ **Intro:** 8,10
         12. $(R(a,b) \wedge R(b,c)) \rightarrow R(a,c)$ — $\forall$ **Elim:** 1
         13. $R(a,c)$ — $\rightarrow$ **Elim:** 11,12,
      14. $\forall j \; R(a,j)$ — $\forall$ **Intro:** 9–13
   15. $\forall j \; R(a,j)$ — $\exists$ **Elim:** 5, 7–14
   16. $\exists i \forall j \; R(i,j)$ — $\exists$ **Intro:** 15
17. $\exists i \forall j \; R(i,j)$ — $\exists$ **Elim:** 4, 6–16
18. $\exists j \forall k \; R(k,j)$ — $\wedge$ **Elim:** 3
19. $\exists k \forall i \; R(i,k)$ — $\wedge$ **Elim:** 2
    20. [c] $\forall k \; R(k,c)$
        21. [d] $\forall i \; R(i,d)$
        22. $R(d,c)$ — $\forall$ **Elim:** 20
            23. [e]
            24. $R(e,d)$ — $\forall$ **Elim:** 21
            25. $R(e,d) \wedge R(d,c)$ — $\wedge$ **Intro:** 22,24
            26. $(R(e,d) \wedge R(d,c)) \rightarrow R(e,c)$ — $\forall$ **Elim:** 1
            27. $R(e,c)$ — $\rightarrow$ **Elim:** 25,26,
        28. $\forall i \; R(i,c)$ — $\forall$ **Intro:** 23–27
    29. $\forall i \; R(i,c)$ — $\exists$ **Elim:** 18, 21–28
    30. $\exists j \forall i \; R(i,j)$ — $\exists$ **Intro:** 29
31. $\exists j \forall i \; R(i,j)$ — $\exists$ **Elim:** 19, 20–30
32. $\exists i \forall j \; R(i,j) \wedge \exists j \forall i \; R(i,j)$ — $\wedge$ **Intro:** 17, 31

Figure 3.8: A transitive relation induces transitive partial relationships..

In the scenario provided by the Figure 3.7b, the granule g_u is partially north of the granule g_v (g_u $north^P$ g_v) and the granule g_v is partially north of the granule g_w (g_v $north^P$ g_w). Therefore, the granule g_u is partially north of the granule g_w (g_u $north^P$ g_w).

Regarding the weak relationship, similar conclusions cannot be made. Given transitive relation R on the domain D, there are circumstances where the weak relationship does not hold the transitivity. One example is provided in Figure 3.7c. The granule g_u is weakly north of the granule g_v (g_u $north^W$ g_v), once there is at least one element in g_u north of all elements of g_v. The granule g_v is weakly north of the granule g_w (g_v $north^W$ g_w), once there is at least one element in g_w where all elements in g_v are north. Still, the granule g_u is not weakly north of g_w.

Finally, the existential relation may also not be transitive in spite of a transitive relation R on the domain D. An example of that is displayed in Figure 3.7d. Although the granule g_u is existentially north of g_v and the granule g_v is existentially north of g_w, the granules g_u and g_w are not existentially related through the relation north of.

The induced relations can be used to specify what kind of properties we intend for certain granularities. For example, the temporal granularities defined by Bettini et al. 2000 are, under our theory of granularities (ToG), granularities defined over the time domain where their granules are related by complete relationships ($<^C$), where the relation ($<$) is induced from the total order verified by the elements of the time domain.

3.1.2 Distance Functions between Granules

Data Mining activity plays an important role on the extraction of patterns that are hidden in very large data sets Committee et al. 2013. Distance/dissimilarity functions are frequently embedded into data mining approaches like clustering, classification, and nearest neighbours search. Instead of having those approaches working on the original domains, it can be advantageous if they work based on the granularities defined for such domains Camossi et al. 2008.

Suppose that there is a granularity G defined over a domain $D = (DS, RS)$, and a real-value distance function d, which quantifies the distance between elements belonging to DS such that $d : DS \times DS \rightarrow R$. Additionally, g_i and g_j denote two granules belonging to G.

The distances between granules can be defined based on the distances of their elements in DS. Here, we consider the following induced distances:

Inner Distance : $d^I(g_i, g_j) = min_{x_i \in E(g_i)} min_{x_j \in E(g_j)} d(x_i, x_j)$

Outer Distance : $d^I(g_i, g_j) = max_{x_i \in E(g_i)} max_{x_j \in E(g_j)} d(x_i, x_j)$

Left Distance : $d^I(g_i, g_j) = max_{x_i \in E(g_i)} min_{x_j \in E(g_j)} d(x_i, x_j)$

Right Distance :$d^I(g_i, g_j) = min_{x_i \in E(g_i)} max_{x_j \in E(g_j)} d(x_i, x_j)$

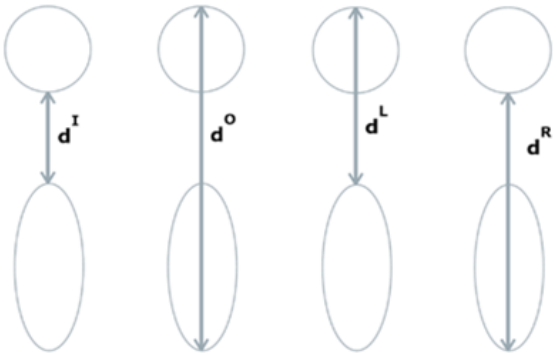


Figure 3.9: Set of induced distances.

The inner distance corresponds to the minimum distance between two granules while the outer distance is the maximum distance. Moreover, the left distance corresponds to the Hausdorff distance from g_i to g_j while the right distance corresponds to the Hausdorff distance (Atallah 1983) from g_j to g_i. The distances are illustrated in Figure 3.9.

Besides the induced distances introduced, several other distances can be defined like the distance between the granules centers of gravity, the minimum between the inner and the outer distance, and so on.

3.2 Relationships between Granularities

Remember that the relationships between granularities allow us to relate different granularities, useful to hold spatiotemporal data at different LoDs.

Two granularities G and H can be related in different manners. In the first place, there are two relations that come naturally from the set theory. Firstly, G and H are equal ($G = H$) if and only if they have precisely the same elements. Note that, a granule is equal to another one if the extents of granules are equal as well as their index values. Furthermore, G is a subset of H ($G \subseteq H$) if and only if for each granule of G there is an equal granule in H.

Others relations can be verified between granularities. In this section, we revisit the majority of the relationships introduced in the literature according to the proposed granularity definition (Bettini et al. 2000; Belussi et al. 2009; Pozzani and Zimányi 2012). For the sake of simplification, in the following formal definitions, we refer to a granule of a granularity by using the lower case letter of the corresponding letter of the granularity. For instance, *each granule's extent of H* can be stated as *each h's extent*.

Two granularities G and H can be related as follows. To complement the discussion, a diagram that illustrates each relation is provided in Figure A.6.

G **is covered by** H $(G \widehat{\sqsubseteq} H)$: the extent of G is contained in the extent of H, formally defined as: $Ext(G) \subseteq Ext(H)$.

G **groups into** H $(G \trianglelefteq H)$: each h's extent is equal to the union of a set of g's extent. The formal definition is: $\forall h \in H, \exists G' \subseteq G : \cup_{g' \in G'} E(g') = E(h)$. However, there may be g's extents that are not contained by any h's extent. From this definition, it can be concluded that G is covered by H: $G \widehat{\sqsubseteq} H$.

G **finer than** H $(G \preceq H)$: each g's extent is contained in one h's extent. The formal definition is: $\forall g \in G, \exists h \in H : E(g) \subseteq E(h)$. There may be h's extents that do not contain some g's extents. From this definition, it can be concluded that G is covered by H: $G \widehat{\sqsubseteq} H$.

G **partitions** H $(G \oplus H)$: each g's extent is contained in one h's extent and each h's extent is equal to the union of a set of g's extent. The formal definition is: $\forall h \in H, \exists G' \subseteq G : \cup_{g' \in G'} E(g') = E(h) \wedge \forall g \in G, \exists h \in H : E(g) \subseteq E(h)$. From this definition, it can be taken that G's extent is equal to the H's extent: $Ext(G) = Ext(H)$, i.e., $G \widehat{\sqsubseteq} H$. Also, $G \preceq H$) and $G \trianglelefteq H$).

G **is a sub-granularity** H $(G \sqsubseteq H)$:for each g's extent there is an equal h's extent. The formal definition is $\forall g \in G, \exists h \in H : E(g) = E(h)$. There may be h's extents inexistent in G. From this definition, it can be concluded that G is covered by H:$G \widehat{\sqsubseteq} H$. This relation is different from G is a subset of H, once we are relating just the extents of granules and not the granules themselves.

The relationships groups into, finer than, sub-granularity, partitions, and the covered (equivalent to image covered in the literature) are relationships proposed and defined accordingly to the granularities definitions (Bettini et al. 2000; Belussi et al. 2009) (see Section 2.2). In this work, we introduce a new relation between granularities labeled as equivalent and formalized as follows.

G **is equivalent to** H $(G \equiv H)$: for each g's extent there is an equal h's extent and for each h's extent there is an equal g's extent. The formal definition is: $\forall g \in G, \exists h \in H : E(g) = E(h) \wedge \forall h \in H, \exists g \in G : E(h) = E(G)$. This relation is different from G is equal to H, once we are relating just the extents of granules and not the granules themselves. From this definition, it can be seen that G's extent is equal to the H's extent: $Ext(G) = Ext(H)$, i.e., $G \widehat{\sqsubseteq} H$. Also, $G \sqsubseteq H$) and $H \sqsubseteq G$).

Through the equivalent relationship, we intend a relationship capable of relating different granularities containing granules with equal extent. For example, we can have two spatial granularities where each granule corresponds to a country. One granularity can be indexing the granules using native names and the other English names.

Let's take a look at the relationships between granularities defined in the context of our examples. Regarding temporal granularities, the granularity $Minutes$ partitions the granularity $Hours$ as well as the granularity $Hours$ partitions the granularity $Days$. The granularity $DaysSubUnits$ has no relationship with the previous mentioned granularities

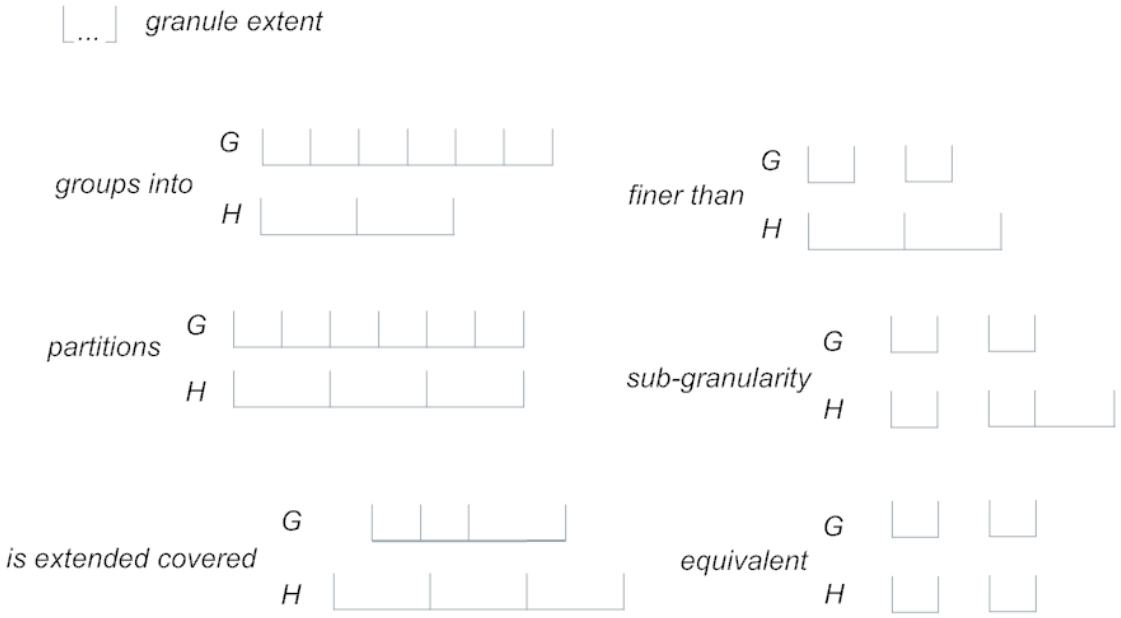


Figure 3.10: Illustration of relationships between granularities.

apart from being possibly covered by the granularities *Hours* and *Minutes*. Additionally, the granularity *WorkDays* where each granule refers to a business day could have been considered. This last granularity is a sub-granularity of the granularity *Days*. Concerning the spatial granularities, the granularity *Parishes* partitions the granularity *Districts*. However, the granularity *ProtectedAreas* has no relation with the previous ones. Another spatial granularity that could have been considered is the granularity *Cities* where each granule corresponds to a city's urban area. This granularity is finer than *Districts*, for instance.

The granularities are instruments to create different "lexicons" to be used to describe realities, but we mainly intend describe spatiotemporal phenomena. Roughly, granularities that share the whole or part of their extension allow us to describe the same reality using different "words". Consequently, when different granularities are related by the relations partitions, finer-than, and groups into, it is possible to describe the same reality with different LoDs.

3.3 Open Issues

Like humans are constantly using granularities in unconscious ways, they also build granularities based on another ones. For example, in Portugal, a district is composed by a set of counties, and a county is a set of parishes. Thus, the granularity *Districts* could have been defined over the granularity *Counties* which in turn could have been defined over the granularity *Parishes*, instead of defining such granularities over $\mathbb{R}^2$.

However, granularities defined over others granularities are not being considered by

the ToG. When a granularity G is defined based on another one, a granule of G will be composed by a set of granules. For example, let's consider the granularity *Districts* created based on *Parishes*, and the granularity *Parishes* defined over $\mathbb{R}^2$. In this case, the extent of a particular district will be a set of granules from *Parishes*, and the extent of particular county will be a "portion" of $\mathbb{R}^2$. Thus, the extents of the granules in *Districts* and *Parishes* are not comparable. This raises an issue because the relationships defined previously like the *finer-than* are no longer applicable, for instance.

On the other hand, granularities may change over time. In an evolution of granularity new granules can emerge, others disappear, and others be split. Even so, some granules keep unchanged. Likely, the evolution of granularities leads to granularities that are different from the ones that are valid in preceding temporal granules. For example, consider the granularity *Parishes* where each granule refers to a parish in Portugal. This Portugal administrative level was recently reorganized. In 2013, some parishes were extinct, others were merged, and others remained unchanged. In order to represent the changes that occurred in *Parishes*, there is a need for the concept of evolution.

Some proposals for the evolution of spatial granularities were found in the literature (Belussi et al. 2009; Pozzani and Zimányi 2012). However, the concept of evolution of a granularity applicable to any domain was not found.

Future work is directed on both issues. On one hand, the definition of granularities over others granularities sharing the same domain of reference should be handled. On the other hand, work on the concept of evolution of granularity can also be done.

3.4 Related Works and their Limitations

This Chapter presents a theory of granularities (ToG) that supports granularities defined over any domain covering the definitions proposed in the literature (Bettini et al. 2000; Wang and Liu 2004; Camossi et al. 2006; Belussi et al. 2009; Pozzani and Zimányi 2012), which focus on a particular domain (temporal or/and spatial) apart from Keet's theory (Keet 2008).

Furthermore, the ToG proposed introduces four induced relations in order to transpose the relations defined in the domains of reference for granules belonging to granularities. None of the works discussed in the literature are capable of such. Some of those relations (that are defined in the original domain) hold properties like symmetric, transitive, reflexive, antisymmetric, and antireflexive. We investigated the circumstances in which the induced relations inherit the properties of the relation defined in the domain of reference. In this study, formal demonstrations were conducted in the natural deduction system.

The ability to transpose the relations defined in the domains for the granules can play a crucial role for the analytical contexts. The establishment of qualitative relations among what happens in space and time is a common practice. For example, we may be interested in whether two events took place at the same time or whether one event took place before

the other, or whether two events overlapped in space (see Section 2.1.1). But if we have described phenomena using granules without having the induced relations, we would have lost the ability to establish those kinds of relationships. This happens because temporal qualitative (Vilain 1982; Allen 1983) relations like before, overlap, during or spatial topological relations (Egenhofer and Sharma 1993; Schneider and Behr 2006) like contains, disjoint are defined over the original domains (time domain and spatial domain, correspondingly). This issue has been ignored by the literature apart from (Bettini et al. 2000) but it lacks generality as they just consider temporal granularities, and the corresponding qualitative temporal reasoning, which is not enough for spatiotemporal phenomena. As opposed to that, in this PhD thesis, we allow relationships to be transposed to granules in any domain of reference.

Another two features of the ToG were devised considering the analytical contexts. One can annotate a granularity with the induced relations as well as with other relations that are important for users' analyses, something that is not considered in the literature. Furthermore, in Data Mining techniques, the usage of distance functions is common. In order to account for this need, we also proposed four induced distance functions. Another subject that has been ignored by the literature.

Therefore, the ToG proposed not only permits to describe a phenomenon but also to reason about it at different LoDs.

Chapter 4

Granularities-based Model

Using the ToG proposed one can represent data, using granularities defined in different domains of reference. This is particularly useful for representing spatiotemporal events at different LoDs as they encompass features with different domains of reference.

Let's consider the dataset of spatiotemporal events about storms occurred in the USA[1] to be our running example of this Chapter. For each storm (i.e., event), we consider the following information: $Space$, $Time$, $Victims$, $Type$ where $Space$ describes the spatial location of the storm (latitude and longitude coordinates), $Time$ specifies the time when the storm occurred (in minutes), $Victims$ describes the number of injured individuals and $Type$ describes the type of storm.

Storms events can be described using different granularities instead of the ones embedded in the domains of reference given by the data provider. To do that, let's define the following granularities based on the ToG proposed.

The granularity $CoordsSeven$ is defined over the two-dimensional space where each granule represents a coordinate with seven decimal cases. The granularity $Raster(0.5km^2)$ is defined over the two-dimensional space where each granule represents a square area of size $0.5km^2$; similarly, consider the granularity $Raster(2km^2)$. Also, consider the granularities $Counties$ and $States$. The granularity $Minutes$, $Hours$, $Days$ are defined over the time domain where each granule represents a minute, an hour, a day, respectively; the granularity $NaturalNumbers$ is defined over $\mathbb{N}$ where each granule corresponds to an element of the corresponding domain. Finally, $StormTypes$ is defined over the domain of type of storms that are considered by the data provider (i.e., tornado, hail, thunderstorm, among others) and each granule corresponds to an element of the corresponding domain.

Some examples of storm events' description using the different granularities are given: "*a large hail occurred on July 8th, 2016 (Days granularity) at Tennessee (States granularity)*

[1] Data available in: `https://www.ncdc.noaa.gov/stormevents/details.jsp`

with zero victims to report"; "*a tornado occurred on May 9th, 2015 15h pm (Hours Granularity) that moved through the Eastland county (Counties granularity) resulting two victims*" or "*a lightning hit Florida in August 12th, 2015 leading to the hospitalization of twenty people.*". For the sake of simplification, the granularities used regarding the non-spatial and temporal attributes of the events were not highlighted. Plus, the granules in each statement were underlined.

The mentioned examples aim to show that one can express individually spatiotemporal events based on the ToG proposed. But not even that is entirely true. In fact, granules are being compliant with abstract real-world entities like a state, a county, an hour, and so on. But there is no theoretical basis framing those granules into a description that something occurred in a phenomenon.

A granular computing approach is proposed to model spatiotemporal phenomena at multiple LoDs labeled as **the granularities-based model**. This approach models a phenomenon through statements rather than just using granules to model abstract real-world entities. Statements are made at some LoD, a concept formally defined which is a key contribution of this work. Based on it, the granularities-based model follows an automated approach to generalize a phenomenon from one LoD to a coarser one. Before we dive into formalisms, let's illustrate the key ideas of the granularities-based model.

A granularities-based model is composed by statements where each one describes something that occurred in a phenomenon. Roughly speaking, granules are used in the statements' arguments. For example, we can model a thunderstorm event through the statement: $storm(Oakland,\ 03/01/2015\ 18h,\ 1,\ thunderstorm)$ where the granules used come from the following granularities $Counties, Hours, Natural\ Numbers, Storm\ Types$.

The thunderstorm event did not occur in the entire extent of Oakland and did not occur from 18h until 18:59. Instead, the thunderstorm occurred in some part of the Oakland county at some point between 18h and 18:59. This form granules' interpretation is called the weak interpretation of granules, as opposed to the strong interpretation (Bravo and Rodríguez 2014). In the latter, the thunderstorm event would have been interpreted as been occurred in the entire extent of Oakland, and would have happened from 18h until 18:59. Therefore, a key property of the granularities-based model is the weak interpretation of granules (Bravo and Rodríguez 2014).

Furthermore, statements can be generalized to coarser LoDs automatically. This occurs, and again roughly speaking, based on the relationship coarsening that occurs between granules: a granule g_1 is a coarsening of another granule g_2 if the extent of g_1 contains the extent of g_2, i.e., $E(g_2) \subseteq E(g_1)$. For example, the previous thunderstorm event can be generalized to $storm(California,\ 03/01/2015,\ 1,\ thunderstorm)$ where the granules used come from the following granularities $States, Days, Natural\ Numbers, Storm\ Types$. Notice that, the granule $California$ is coarsening of the granule $Oakland$, the granule $03/01/2015\ 18h$ is coarsening of the granule 03/01/2015, the granule 1 is coarsening of the granule 1 and the granule $thunderstorm$ is coarsening of the granule $thunderstorm$.

The expression *roughly speaking* was used to mentioned that granules are used in the statements' arguments. This was done in order to keep the presentation of the core ideas of the granularities-based model simple. However, the concept called the *granular term* is proposed that is the basis for what is used in the statements' arguments. The motivation behind this concept along with its formal definition is given below.

4.1 Granular Terms

The granules result from defining a granularity over a data domain. These may or may not match abstract real world entities. For example, in space, the granularity $Countries$ contains granules compliant with entities like Portugal, USA, among others. But the $Raster(2km2)$ granularity does not match any particular entity, that is, the granules represent only fixed-length cells of size: two square kilometers. In time, for example, the granularity *Days* contains granules that correspond to entities such as December 1, 1987, or February 17, 1988. But we could have defined a granularity where granules are not compliant with any particular concept.

Granularities might contain granules representing some concept/entity. Yet, in several scenarios, it is desirable to use granules from one granularity to express particular concepts/entities, at such granularity, which are not captured by the granules themselves.

Looking at time, common temporal concepts (i.e., time primitives) are time instant or time interval (see Section 2.1.1). Consider the granularity of *Days* displayed in Figure 4.1, illustrating twelve days. In the same way, a time interval is defined over the time domain, we should also be able to define a time interval at the granularity of *Days* as displayed in Figure 4.1. Actually, it's something that humans do unconsciously like someone mentioned that a wildfire consumed forest from August 17, 2015 to August 19, 2015. In this particular case, that someone uses a discourse at the granularity of *Days* and not at the time domain.

Likewise, common spatial concepts like point, line, regions were presented in Section 2.1.1. Therefore, and independently from the data domain, one might want to represent a particular concept recurring to granules belonging to a granularity instead of their domains of reference.

But how do we express time primitives in terms of a temporal granularity, for instance? How these time primitives are transposed to granular computing? Time primitives have been defined over the time domain (Vilain 1982; Allen 1983). Likewise, the spatial primitives have been defined over two-dimensional or three-dimensional space (Ryden 2005), and not in terms of spatial granularities.

To meet this need, we introduce the granular term concept. Granular terms are built based on function symbols and granules from a single granularity. . Let f be an n-ary or a variadic function symbol and G a granularity. A n-ary function symbol has a fixed arity while a variadic function symbol takes a variable number of arguments. A granular term is $f(g_1, \ldots, g_n)$ such that $g_i \in G$ for all $1 \leq i \leq n$; or, a granular term is $f(t_1, \ldots, t_n)$ such that t_i

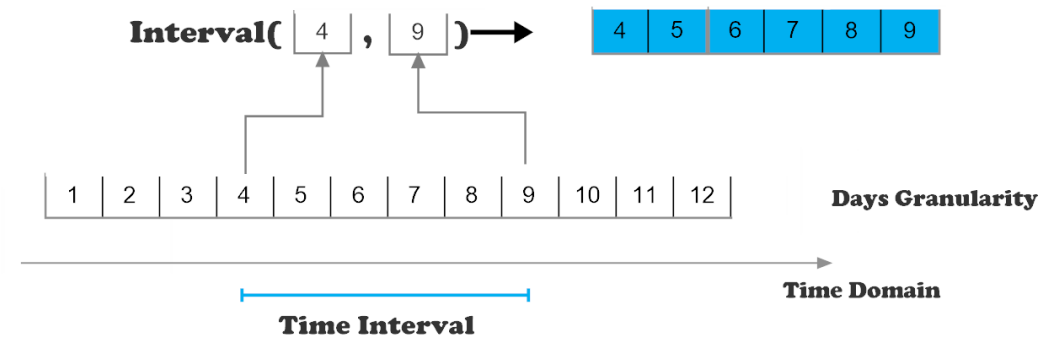


Figure 4.1: An illustration of a time interval be defined in terms of a temporal granularity.

is a granular term defined using the granularity G for all $1 \leq i \leq n$. Granular terms in the form of $f(g_1, \ldots, g_n)$ are simple, and the ones in the form of $f(t_1, \ldots, t_m)$ are compound. Finally, granular terms can also be built using the identity function symbol $Id(g \in G)$, which is useful to use granules that already represent a particular concept.

$Interval$(03/01/2015 $18h$, 03/01/2015 $19h$) is an example of a simple granular term using granules from $Hours$; $MultiInterval(Interval$(03/01/2015 $18h$, 03/01/2015 $19h$), $Interval$(03/01/2015 $21h$, 03/01/2015 $22h$)) is an example of a compound granular term using granules from $Hours$; and, $Id(Oakland)$ is an example of a granular term built based on the identity function symbol and the granularity $Counties$.

A function symbol allows building granular terms by using a collection of granules that represents a particular concept. As such, each function symbol contains its own signature establishing the needed restrictions to build granular terms of type f. For example, the $Interval$ function symbol needs to establish additional constraints in order to disallow improper granular terms of $Interval$ like the ones in the form of $Interval(Interval(a,b), Instant(c))$.

This work formalizes the following function symbols: $Instant$ and $Interval$ (see Section 4.3.1), and $Cell$ and $RasterRegion$ (see Section 4.3.2). These allow modeling time instants, time intervals, cells or raster regions, respectively.

Other examples of function symbols can be pointed out but we left their formalization for future work. For example, the need to represent spatial features in vector space like points, lines, polygons, and a set of polygons, among others examples is common. Also, in several applications scenarios, the concept of trajectory is crucial to model the trajectories made by people, cars, animals, among others. The function symbols needed depend on the phenomenon under study and the underlying data type to record it (see Section 2.1.1).

Granular terms are used in the statements' arguments. Therefore, the following event: "*a tornado occurred on May 9th, 2015 between 15:32 and 15:52 pm that moved through two cells of size* $2km^2$ *at Eastland county resulting in two victims*" can be modeled like $storm(RasterRegion(cell_1, cell_2), Interval$(09/05/2015 15 : 45, 09/05/2015 15 : 52), Id(20), $Id(tornado))$ where the granules would come from

$(space, Raster(2km^2))$, $(time, Minutes)$, $(victims, Natural\ Numbers)$, $(type, Storm\ Types)$.

4.2 Predicate and Atoms

A phenomenon is modeled through a collection of statements. These are built based on a definition of a predicate. A predicate P contains a set of arguments $Args(P)$, and its signature declares for each one of its arguments a set of granularities $G_{(P,arg)}$ and function symbols $F_{(P,arg)}$ that can be used. Let $granularTerm(G, f)$ denote a granular term of f using granules from the granularity G. This way, a well-formed atom (i.e., a statement) is in the form of $P(\tau)$ with $\tau = \{(arg,\ granularTerm(G,\ f)) \mid arg \in Args(P) \wedge f \in F_{(P,\ arg)}\} \wedge G \in G_{(P,\ arg)}$. τ denotes the tuple of terms of an atom.

Let's introduce the *storm* predicate in order to model storm events. For each argument of the *storm* predicate, we declare the following set of valid granularities:

- $G_{(storm,space)} = \left\{CoordsSeven, Raster(0.5km^2), Raster(2km^2), Counties, States\right\}$
- $G_{(storm,time)} = \{Minutes, Hours, Days\}$
- $G_{(storm,victims)} = \{Natural\ Numbers\}$
- $G_{(storm,type)} = \{Storm\ Types\}$

Also, for each argument, we declare the following set of valid function symbols:

- $F_{(storm,space)} = \{Cell, RasterRegion, Id\}$;
- $F_{(storm,time)} = \{Instant, Interval\}$
- $F_{(storm,victims)} = \{Id\}$
- $F_{(storm,type)} = \{Id\}$

A well-formed atom of the *storm* predicate uses granules from the valid granularities declared for each argument as well as the valid function symbols. For example, the atom $o_1 = storm\left(\left\{Cell\left(cell_1^{0.5km^2}\right), Interval(10/5/2014\ 16:40,\ 10/5/2014\ 16:45), Id(2), Id(tornado)\right\}\right)$ describes a tornado occurred on May 10th, 2014 between 16:40 and 16:45 which moved inside an area of $0.5km^2$ and resulted in 2 victims.

We assume that there is **a base granularity** for each set of valid granularities of each argument of a predicate P. A base granularity of an argument of a predicate P is a granularity that is related with any other granularity valid on such argument through the relation *finer than*. A base granularity in $G_{(P,\ arg)}$ is formally defined as follows: $\exists! G_{base} \in G_{(P,\ arg)} :\ G_{base} \preccurlyeq G \in G_{(P,\ arg)}$. For example, the base granularity in $G_{(storm,\ space)}$ is $CoordsSeven$ and the base granularity in $G_{(storm,\ time)}$ is $Minutes$.

An atom describes something that happens in a spatiotemporal phenomenon. The set of granularities involved in an atom defines the LoD at which something is described. For example, the LoD of the atom o_1 is $LoD(o_1) = \{(space, Raster(0.5km^2), (time, Minutes), (victims, Natural\ Numbers), (type, StormsTypes)\}$. We define the valid LoDs of a predicate as follows.

Let $\gamma = P\Big(\Big\{\Big(arg_1, granular\ term_1\Big), \ldots, \Big(arg_n, granular\ term_n\Big)\Big\}\Big)$ be an atom; the set $\{(arg_1, G_1), \ldots, (arg_n, G_n)\}$ describing the granularity used in each argument defines its $LoD(\gamma)$.

Definition 4.1 (Valid Levels of Detail of a Predicate)**.** Let P be n-ary predicate and its signature $P\Big(\Big\{\Big(arg, \big(G_{(P,\ arg)}, F_{(P,\ arg)}\big)\Big) \mid arg \in Args(P)\Big\}\Big)$ defining a set of valid granularities and function symbols for each argument; then $\mathscr{L}^P = \bigotimes_{\text{arg} \in Args(P)} G_{(P,\ arg)}$ is the set of valid LoDs of the predicate P.

The set of valid LoDs of a predicate results from the Cartesian product among the sets of valid granularities. In our example $\mathscr{L}^{storm} = G_{(storm,space)} \times G_{(storm,time)} \times G_{(storm,victims)} \times G_{(storm,type)}$. Therefore, the number of LoDs in $\mathscr{L}^{storm}$ that one can observed the phenomenon is 15. This means that, in our simple example, one user might need to go through 15 LoDs in order to understand in what of LoDs some patterns are better perceived without probably really knowing what patterns might be in the data. This paradox is what we aim to solve in this PhD thesis, providing an overview of potential patterns that might be in the data, and simultaneously, telling what in LoDs are suitable to study them.

Two valid LoDs α and β of a predicate P can be related through a relationship called *more detailed than*. We introduce the *more detailed than* relation between LoDs.

Definition 4.2 (α is more detailed than β)**.** Let P be n-ary predicate; let $\alpha \in \mathscr{L}^P$ and $\beta \in \mathscr{L}^P$ be two valid LoDs of P such that $\alpha = \{(arg_1, G_1), \ldots, (arg_n, G_n)\}$ and $\beta = \{(arg_1, H_1), \ldots, (arg_n, H_n)\}$; α is more detailed than β, $\alpha \preccurlyeq_L \beta$, if and only if, $G_i \preccurlyeq H_i$ for all $1 \leq i \leq n$.

The set of all valid LoDs $\mathscr{L}^P$ of a predicate P with the relation *is more detailed than* ($\preccurlyeq_L$) define a poset: $\mathbb{L}^P = (\mathscr{L}^P, \preccurlyeq_L)$. There is only one least LoD α in $\mathbb{L}^P$ such that for every LoD β in $\mathbb{L}^P$, $\alpha \preccurlyeq_L \beta$. Note that, the least LoD of a predicate P is composed by the set of base granularities of the corresponding arguments, which we denote by the base LoD of P. In our example, part of the Hasse diagram regarding the poset $\mathbb{L}^{storm}$ is illustrated in Figure 4.2, in which the base LoD of the *storm* predicate corresponds to the $LoD_0 = (space, CoordsSeven), (time, Minutes), (victims, Natural\ Numbers), (type, Storm\ Types)$.

In order to have atoms at multiple LoDs, we propose to take an atom in one LoD and express it at a coarser one. To that end, each function symbol must have associated a set of generalization rules $\mathbb{G}_f$, allowing each argument to have its own process of generalization.

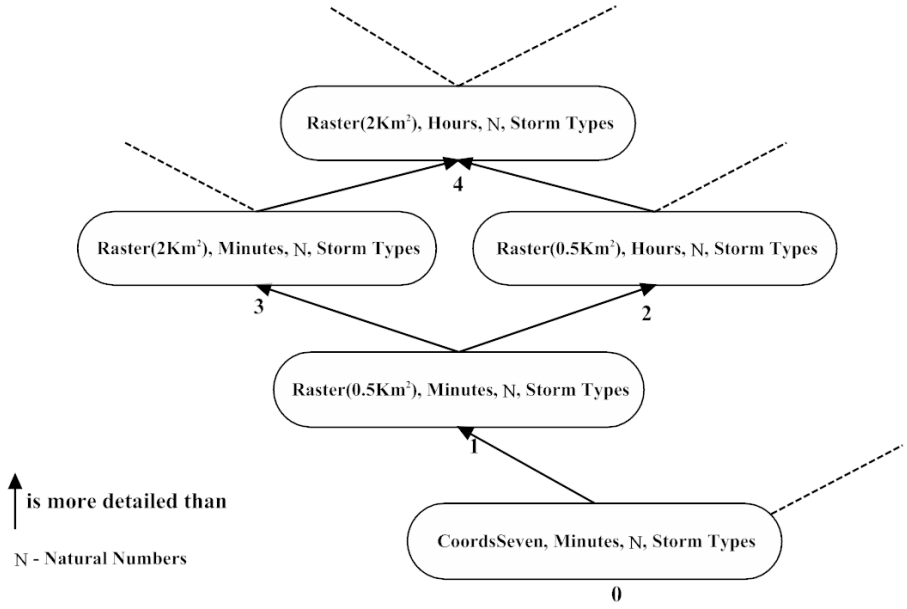


Figure 4.2: Part of the Hasse diagram concerning the poset $\mathbb{L}^{storm}$.

This way, the generalization can turn a time interval into a time instant, simplify a raster region, or even turn a raster region into a cell (i.e., generalization-reduction process as detailed in Section 2.3.2). The formalization of the generation rules concerning the function symbols $Instant$, $Interval$, $Cell$ and $RasterRegion$ are detailed ahead in Section 4.3.

A simple example of generalization is the case of the identity function symbol. One granular term $Id(g_1 \in G)$ is generalized into another $Id(g_2 \in H)$ if the extent of g_2 contains the extent of g_1, i.e., $E(g_1) \subseteq E(g_2)$ and G is finer than H ($G \preccurlyeq H$.

The generalization of atoms occurs between two valid LoDs α and β of a predicate P such that α is more detailed than β ($\alpha \preccurlyeq_L \beta$). This way, each atom is generalized from α to β by applying the generalization rules to each granular term specified in each argument of an atom.

Let's consider the atoms a_1,..., a_4, shown in Figure 4.3, expressed at the LoD_1 of the storm predicate. The atoms are describing the spatial location of lightnings using the granularity of $Raster(0.5km^2)$ and when it occurred with the granularity of $Minutes$. Consider that it would be desirable to describe the locations of lightnings and when they occurred with coarser granularities ($Hour$ and $Raster(2km^2)$).

Making the generalization of atoms, we can produce a set of atoms at the valid LoD_4 based on the set of atoms at the LoD_1 of the storm predicate. For the sake of simplification, we are not going into detail regarding the generalization rules as they will be detailed in Section 4.3. In this example, the granular terms are just generalized based on the coarsening relation that occurs between granules. Informally, the atom a_4 can be generalized into the atom a_8 once the extent of the granule $cell_1$ is contained by the extent of the granule

$cell_a$ - $E(cell_1) \subseteq E(cell_a)$; the extent of the granule $02-03-2013\ 18:35$ is contained by the extent of the granule $02-03-2013\ 18$ - $E(02-03-2013\ 18:35) \subseteq E(02-03-2013\ 18)$; and, similarly, $E(1) \subseteq E(1)$, $E(Lighting) \subseteq E(Lighting)$. Similarly, the atom generalization is applied for the remaining atoms at the LoD_1.

Atoms at the LoD 4 of the storm predicate

a_8 = storm((space, Cell($cell_a$)), (time, Instant(01-01-2012 07h am)), (victims, one), (type, Lightning))

a_7 = storm((space, Cell($cell_a$)), (time, Instant(01-01-2012 07h am)), (victims, one), (type, Lightning))

a_6 = storm((space, Cell($cell_a$)), (time, Instant(02-03-2013 18h pm)), (victims, one), (type, Lightning))

a_5 = storm((space, Cell($cell_a$)), (time, Instant(02-03-2013 18h pm)), (victims, two), (type, Lightning))

Atoms at the LoD 1 of the storm predicate

a_4 = storm((space, Cell($cell_1$)), (time, Instant(01-01-2012 07:45 am)), (victims, one), (type, Lightning))

a_3 = storm((space, Cell($cell_2$)), (time, Instant(01-01-2012 07:10 am)), (victims, one), (type, Lightning))

a_2 = storm((space, Cell($cell_3$)), (time, Instant(02-03-2013 18:15 pm)), (victims, one), (type, Lightning))

a_1 = storm((space, Cell($cell_4$)), (time, Instant(02-03-2013 18:35 pm)), (victims, two), (type, Lightning))

Figure 4.3: Example of atoms at different valid LoDs of the storm predicate.

The atom generalization provides an instrument to automatically generalize a phenomenon for coarser LoDs. A granularities-based model may contain equal atoms, i.e., atoms composed by the same granular terms in some valid LoD of a predicate P in spite of the fact that they are referring to different occurences in a phenomenon. As can be seen in Figure 4.3, the atoms at the valid LoD_1 of the storm predicate are discernible from each other while at the valid LoD_4 some atoms are equal, namely a_7, a_8. Note that, they are describing distinct occurrences of lightnings.

In general, at a valid LoD of a predicate P, there may be atoms equal to each other. Furthermore, as the atoms are described through coarser valid LoDs of P, the number of equal atoms tends to increase. When there are equal atoms at some LoD of a predicate P, we are interested in performing synthesis of atoms in order to reduce the number of atoms that describe a phenomenon. Thereby, we introduce the concept of **granular synthesis**.

Beforehand, and without loss of generality, we assume that any atom of form $P\left(\left\{\left(arg_1,\ granular\ term_1\right), \ldots, \left(arg_n,\ granular\ term_n\right)\right\}\right)$ can be expressed equivalently as $G_{Syn}\left(P\left(\left\{\left(arg_1,\ granular\ term_1\right), \ldots, \left(arg_n,\ granular\ term_n\right)\right\}\right), 1\right)$, where G_{Syn} is a reserved predicate such that the first argument contains an atom of a predicate P and the second one indicates the number of occurrences of such atoms which, in this case, is one.

Definition 4.3 (Granular Synthesis)**.** Let P be n-ary predicate; let $\tau_{\mathbb{A}}$ be a tuple of terms; let $\mathbb{A}$ be a set of atoms at a valid LoD of P such that any atom $a \in \mathbb{A}$ is of form $G_{Syn}(P(\tau_{\mathbb{A}})\ ,\ fr_{\mathbb{A}})$; then, a function $g : \mathbb{A} \longrightarrow G_{Syn}(P(\tau),\ fr)$ produces a granular synthesis where fr is the sum of all frequency values $fr_{\mathbb{A}}$ in $\mathbb{A}$ such that $fr \in \mathbb{N}$.

A granular synthesis makes a summary of a set of equal atoms at a valid LoD of a predicate P. Thus, a granular synthesis is an instrument to reduce the volume of atoms at some LoD of a predicate P. As shown in Figure 4.4, the atoms a_7, a_8 resulted in the granular synthesis a_9. The remaining atoms of the storm predicate are expressed also as granular syntheses in spite of the fact that their count is equal to one.

Atoms at the LoD 4 of the storm predicate (expressed through granular syntheses)

$a_9 = G_{Syn}$(storm((space, Cell(cell$_a$)), (time, Instant(01-01-2012 07h am)), (victims, one), (type, Lightning)), 2)

$a_{10} = G_{Syn}$(storm((space, Cell(cell$_a$)), (time, Instant(02-03-2013 18h pm)), (victims, one), (type, Lightning)), 1)

$a_{11} = G_{Syn}$(storm((space, Cell(cell$_a$)), (time, Instant(02-03-2013 18h pm)), (victims, two), (type, Lightning)), 1)

Atoms at the LoD 4 of the storm predicate

a_8 = storm((space, Cell(cell$_a$)), (time, Instant(01-01-2012 07h am)), (victims, one), (type, Lightning))

a_7 = storm((space, Cell(cell$_a$)), (time, Instant(01-01-2012 07h am)), (victims, one), (type, Lightning))

a_6 = storm((space, Cell(cell$_a$)), (time, Instant(02-03-2013 18h pm)), (victims, one), (type, Lightning))

a_5 = storm((space, Cell(cell$_a$)), (time, Instant(02-03-2013 18h pm)), (victims, two), (type, Lightning))

Figure 4.4: Example of granular syntheses at the LoD_4 of the storm predicate.

The concepts introduced and illustrated lead to a model that allows us to look at a phenomenon and analyze it at different LoDs, formalized as follows.

Definition 4.4 (Granularities-based Model)**.** Let $\mathcal{G} = \{\mathcal{A}(G_1), \ldots, \mathcal{A}(G_n)\}$ be a set of annotated granularities, $\mathcal{P}$ a set of predicates and $\mathcal{F}$ a set of function symbols. Each predicate has defined its signature. A granularities-based model $\mathcal{M}$ is a set of well-formed atoms.

1. $P(\tau)$ with $\tau = \{(arg, granularTerm(G, f)) \mid \text{arg} \in \text{Args(P)} \wedge f \in F_{(P, arg)}\} \wedge G \in G_{(P, arg)}$.
2. $G_{Syn}(P(\tau), \text{fr})$ such that $fr \in \mathbb{N}$.

4.3 Function Symbols

Function symbols need to be formalized in order to allow us to define granular terms. Regardless of the function symbol used to define a granular term, it has to obey the general definition of the term granular. In this work, we formalize the following function symbols: *Instant* and *Interval* (see Section 4.3.1), and *Cell* and *RasterRegion* (see Section 4.3.2). Before we present them, there is a key aspect to be discussed.

The concept of the granular term brings a tremendous advantage that, to the best of our knowledge, has been ignored by granular computing. The ToG proposed introduced four induced relationships that allow to bringing relations defined over the original domain for the granules. For example, in space, we can evaluate whether an element belonging to a two-dimensional space is north of another one. Thus, we can also check if

one granule is completely, partially, weakly or existentially north of another. In time, we can assess if an element belonging to a time domain occurs before another one. Therefore, we can also check whether one granule is completely, partially, weakly or existentially before another.

The establishment of qualitative relations between what happens in space and time is a common practice. For example, we may be interested in whether events overlapped in time or whether two events are overlapped in space (see Section 2.1.1). However, the granules themselves do not embrace any particular concept, and consequently, the induced relations are not enough to bring qualitative reasoning into the granular domains.

By introducing granular terms, we can now have concepts like time interval, time instant, cell, region in a granular domain, and therefore, we can bring qualitative relations, that are defined in the domains of reference, to granular domains. This is accomplished by using the induced relations on top of the granular terms. In this work, this was done for temporal granular terms. This resulted in a supplementary contribution of this PhD thesis.

A formal study about what happens to temporal relations between temporal terms when these are generalized was made, thus allowing, reasoning about temporal relations at different granularities. This study is introduced in Section 4.3.1 but details can be found in Appendix B (Topological Relations on Temporal Granular Terms). This study extends the results obtained by Euzenat and Montanari 2005 obtained in a different line of research as the authors' starting point is a qualitative time representation (Euzenat and Montanari 2005). Euzenat and Montanari 2005 assumed that the generalization of any interval of time results always in an interval of time. However, the generalization of an interval of time might result in an instant of time. In those cases, Euzenat's conversion table is no longer applicable, something that was handled in our study.

4.3.1 Temporal Granular Terms

In order to represent time, we introduce temporal granular terms, which are built using temporal granularities. As to the time domain, time instants have no duration. In contrast, a time interval is the set of all time instants between a starting point and a finishing point.

Let T be a temporal granularity. To represent a time instant of T, we introduce the $Instant$ function symbol defined as follows: $Instant(t)$ where $t \in T$.

Two granules of T can be related through the induced complete relationship $<^C$ (see Section 3.1.1) in order to tell whether a granule of T occurs before another one. In order to represent time intervals of T, we introduce the $Interval$ function symbol.

Definition 4.5 (Time Interval)**.** Let $Interval$ be a function symbol and its arity is equal to two; let t^- and t^+ be granules of T such that $t^- <^C t^+$ (also mentioned as the endpoints of the interval); a time interval of T $\{t_i \in T \mid t^- <^C t_i <^C t^+\}$ is denoted by $Interval(t^-, t^+)$.

Granular terms of *Instant* or *Interval* should be interpreted in the context of the temporal granularity used to build them. For instance, a granule from a granularity *Hours* represents an hour of time and it should not be considered a time interval in the context of this work but rather an indivisible moment of time. Recall that, a granule is a non-decomposable entity. Therefore, granules from a temporal granularity T are interpreted as time instants.

Based on the temporal granular terms presented, we can build atoms describing that something occurred in some time instant or time interval of T. A well-formed atom describing a hail event is for example: $o_2 = storm\left(\left\{Cell\left(cell_1^{0.5km^2}\right), Interval(11/5/2014\,16:40,\ 11/5/2014\,16:45), Id(1), Id(hail)\right\}\right)$. In this example, the interval of time is a granular term defined at granularity *Minutes*.

Allen 1983, Vilain 1982 and point algebras model qualitative relations between time intervals, time intervals and time instants (or vice-versa), and time instants, respectively, which are defined over the time domain. Since we can bring the relations of the domain into the granularities (see Section 3.1.1), we transpose the topological relations for temporal granular terms.

Let a=$Instant(\alpha)$, b=$Instant(\beta)$ be granular terms of T. a can occur before b ($\alpha <^C \beta$), both time instants can be equal ($\alpha = \beta$), or a can occur after b ($\alpha >^C \beta$). On the other hand, let $c = Interval(\alpha^-, \alpha^+)$ and $d = Interval(\beta^-, \beta^+)$ be granular terms of T. c and d can be related as follows (the symmetric relations are not displayed):

1. c before d *iff* $\alpha^+ <^C \beta^-$
2. c equals d *iff* $(\alpha^- = \beta^-) \wedge (\alpha^+ = \beta^+)$
3. c overlaps d *iff* $\left(\alpha^- <^C \beta^-\right) \wedge \left(\alpha^+ >^C \beta^-\right) \wedge (\alpha^+ <^C \beta^+)$
4. c meets d *iff* $\alpha^+ = \beta^-$
5. c starts d *iff* $\alpha^- = \beta^- \wedge \alpha^+ <^C \beta^+$
6. c during d *iff* $\alpha^- >^C \beta^- \wedge \alpha^+ <^C \beta^+$
7. c finishes d *iff* $\alpha^+ = \beta^+ \wedge \alpha^- >^C \beta^-$

Last but not least, let e=$Instant(\alpha)$ and $f = Interval(\beta^-, \beta^+)$ be granular terms of T. e and f can be related as follows (the symmetric relations are not displayed):

1. e before f *iff* $\alpha <^C \beta^-$
2. e starts f *iff* $\alpha = \beta^-$
3. e during f *iff* $\beta^- <^C \alpha <^C \beta^+$
4. e finishes f *iff* $\alpha = \beta^+$

5. e after f *iff* $\beta^{+} <^{C} \alpha$

Generalization rules are defined for each function symbol so that the generalization of atoms can be performed automatically. We define generalization rules applicable to temporal granular terms. When a temporal granular term is generalized, an instant or an interval of time can remain an instant or an interval, correspondingly, but with less precision; or a time interval can become a time instant. The generalization of temporal terms is formalized as follows. Let T_1 and T_2 be temporal granularities such that T_1 is finer than T_2 ($T_1 \preccurlyeq T_2$).

An instant of time a_1=$Instant(\alpha)$ of T_1 can be generalized into an instant of time a_2=$Instant(\alpha')$ of T_2 through

$$\mathbb{G}_{\text{Instant}}: (a_1, T_1) \longrightarrow (a_2, T_2) \text{ if and only if } \exists!\ \alpha' \in T_2:\ \mathrm{E}(\alpha) \subseteq \mathrm{E}(\alpha')$$

that is if there is exactly one granule α' belonging to T_2 such that the extent of α is contained by the extent of α'. For example, the $Instant$(10–5–2014 16 : 40*)* at granularity $Minutes$ is generalized into the $Instant$(10–5–2014 16h) at granularity $Hours$.

An interval of time $a_1 = Interval(\alpha^{-}, \alpha^{+})$ of T_1 can be generalized into an interval of time $a_2 = Interval\left(\alpha'^{-}, \alpha'^{+}\right)$ of T_2 through

$$\mathbb{G}_{\text{Interval}}: (a_1, T_1) \longrightarrow (a_2, T_2) \text{ if and only if } \exists!\ \alpha'^{-} \in T_2:\ \mathrm{E}(\alpha^{-}) \subseteq \mathrm{E}\left(\alpha'^{-}\right) \text{ and } \exists!\ \alpha'^{+} \in T_2:\ \mathrm{E}(\alpha^{+}) \subseteq \mathrm{E}\left(\alpha'^{+}\right)$$

That is if there is exactly one granule α'^{-} belonging to T_2 such that the extent of α^{-} is contained by the extent of α'^{-} and, if there is exactly one granule α'^{+} belonging to T_2 such that the extent of α^{+} is contained by the extent of α'^{+}. Moreover, **an interval of time $a_1 = Interval(\alpha^{-}, \alpha^{+})$ of T_1 can be generalized into an instant of time a_2=$Instant\left(\alpha'\right)$ of T_2** through

$$\mathbb{G}_{\text{Interval}}: (a_1, T_1) \longrightarrow (a_2, T_2) \text{ if and only if } \exists!\ \alpha' \in T_2:\ \mathrm{E}(\alpha^{-}) \subseteq \mathrm{E}\left(\alpha'\right) \wedge \mathrm{E}(\alpha^{+}) \subseteq \mathrm{E}\left(\alpha'\right).$$

That is if there is exactly one granule α' belonging to T_2 such that the extent of α^{-} and α^{+} is contained by the extent of α'. For example, the $Interval$(10–5–2014 16 : 40, 10–5–2014 17 : 45) at granularity $Minutes$ is generalized into $Interval$(10–5–2014 16h, 10 – 5 – 2014 17h) at granularity $Hours$, or into $Instant$(10–5–2014) at granularity $Days$.

The generalization of temporal granular terms may affect the temporal topological relationships held between pairs of temporal granular terms. On one hand, the type of relationship may change. For instance, we might have a relation between two-time intervals that may turn into a relation between a time interval and a time instant. On

the other hand, there are scenarios where the type of topological is kept but the actual relation (e.g., before) is changed (e.g., to equal).

An overview of the possible transitions between types of topological relations is given in Figure 4.5. To each scenario, an example is given based on the temporal granularities T_1 and T_2 illustrated in Figure 4.6. The granules of T_1 are identified by a number and the granules of T_2 by a letter to simplify the discussion that follows.

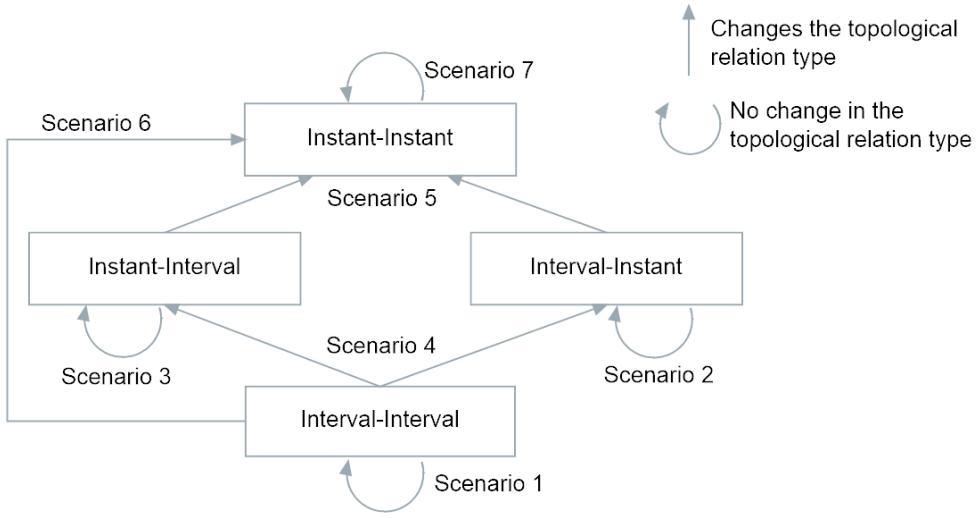


Figure 4.5: Possible transitions in the relationships between pairs of temporal terms.

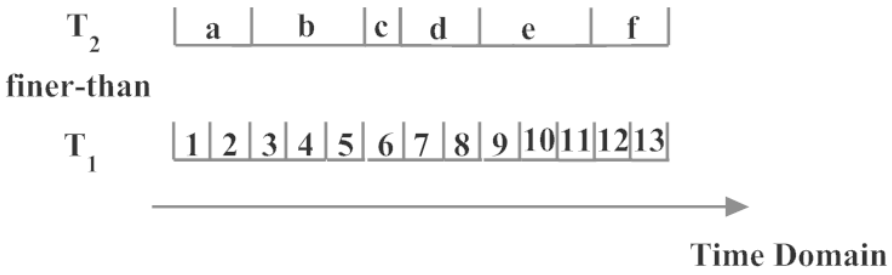


Figure 4.6: Example of two temporal granularities related by the finer-than relationship.

Consider the following granular terms of *Interval* using granules from T_1: α=*Interval*(1, 5) and β=*Interval* (3, 6) such that α *overlaps* β. After the generalization of α and β to the granularity T_2, α and β became α'=*Interval*(a, b) and β'=*Interval*(b, c), respectively. As such, the type of relation is kept (scenario 1) but the actual relation is changed to α' *meets* β' as can be observed in Figure 4.7.

A relation between two-time intervals can turn into a relation between an instant and an interval of time or the other way around (scenario 4). For example, α=*Interval*(1, 2) occurs before than β=*Interval* (3, 6). After their generalization, we get: α' = a occurs before than β' = *Interval*(b, c) as displayed in Figure 4.8.

Also, two-time intervals can turn into a relation between two-time instants (scenario

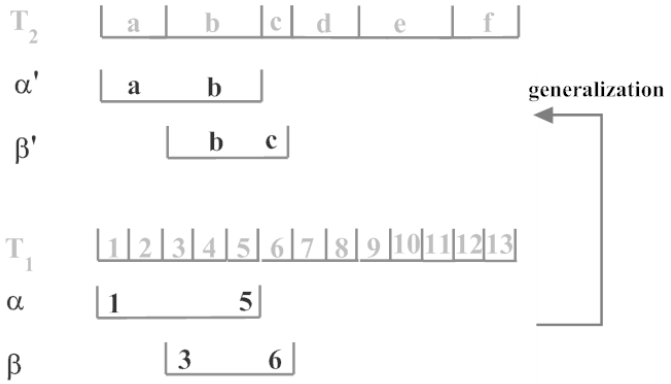


Figure 4.7: First illustration of the generalization of temporal granular terms.

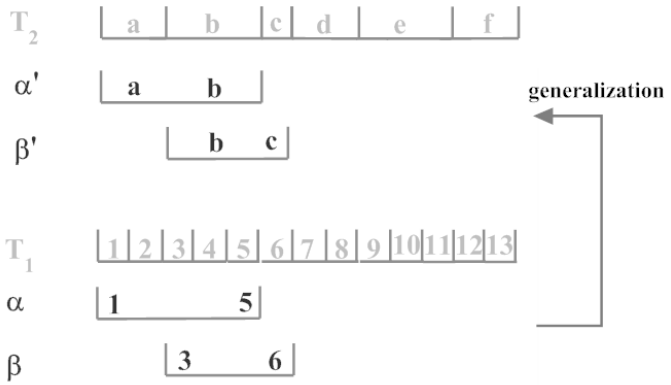


Figure 4.8: Second illustration of the generalization of temporal granular terms.

6). For example, α=$Interval(3,4)$ meets β=$Interval$ (4, 5). After their generalization we get: α' = b equals β' = b. Moreover, a relation between an instant and an interval of time (or vice-versa) can be kept but the actual relation can be changed (scenario 2 or 3). For example, α= 4 occurs during β=$Interval$ (3, 7) turns into α'= b starts β'=$Interval$ (b, d). Furthermore, a relation between an instant and an interval of time (or vice-versa) can turn into a relation between two-time instants (scenario 5). For instance, α= 3 starts β=$Interval$ (3,5) leads to α'= b equals β'= b. Last but not least, a relation between two-time instants may be kept or changed (scenario 7). For example, α= 9 occurs before β= 10 becomes α'= e equals β'= e.

A detailed study of the possible changes in all scenarios, as well as in what conditions they occur is provided in Appendix B (Topological Relations on Temporal Granular Terms).

4.3.2 Spatial Granular Terms

In order to represent spatial features in raster space, we introduce spatial granular terms. These are built based on granularities defined over two-dimensional space where granules have equal square sized extents, i.e., raster granularities.

For the contexts of raster data, points are represented as cells, and raster regions are groups of contiguous cells that portray the shape of an area. Using granules from raster granularities, one may want to use granular terms to describe cells or raster regions.

In general, a region is mentioned as a set of connected cells, i.e., one can "travel" from any cell to any other in the region by following its neighbors. However, there are different definitions of raster regions (Kong and Rosenfeld 1989; Egenhofer and Sharma 1993). These definitions rely on the neighborhood concept. The 4-neighbors of a cell consist in the cells that share the vertical and horizontal sides and the 8-neighbors are the ones sharing diagonal sides in addition to the 4-neighbors.

A region (without holes) is, in general, defined by a Jordan[2] curve which divides a raster space into two parts (interior and exterior). However, if we consider 4-adjacency or 8-adjacency, a paradox emerges in some curves (Kong and Rosenfeld 1989) as displayed in Figure 4.9.

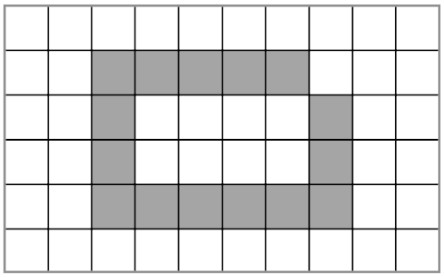


Figure 4.9: An example of a curve in the raster space.

When we consider the 4-adjacency, the curve is not closed and the inside of the curve is not connected to the outside of the curve. This violates the Jordan Curve theorem, once a non-closed curve is dividing space into two parts. On the other hand, if the 8-adjacency is considered, the curve is closed and the inside of the curve is connected to the outside of the curve. This also violates the theorem because a closed curve has not separated space into two parts.

One approach to overcome this problem is to consider different adjacency rules regarding a region and its complement (Kong and Rosenfeld 1989). In this work, we do not aim to propose a new raster region definition and we will adopt the mixed adjacency model (8, 4) to define a raster region, i.e., a raster region is 8-connected and its complement is 4-connected.

Let S be a raster granularity. To represent a cell of S, we introduce the *Cell* function symbol defined as follows: *Cell*(c) where c∈S. In order to represent a raster region of S, we introduce the *RasterRegion* function symbol as follows.

Definition 4.6 (Raster Region)**.** Let *RasterRegion* be a variadic function symbol; let $c_1, \ldots, c_n$ be granules of a granularity S, i.e., $c_i \in S$ for all $1 \leq i \leq n$; a raster region of S

[2]Jordan Curve Theorem: http://mathworld.wolfram.com/JordanCurveTheorem.html

is denoted by *RasterRegion* ($c_1, \ldots, c_n$) where {$c_1, \ldots, c_n$} is a set and their elements are 8-connected, the S\{$c_1, \ldots, c_n$} is 4-connected, and $n > 1$.

Based on the spatial granular terms presented, we can define atoms describing that something occurred in a particular cell or region of S. For example: $o_3 = storm(RasterRegion\left(\text{cell}_1^{0.5\text{km}^2}, \text{cell}_2^{0.5\ \text{km}^2}, \text{cell}_3^{0.5\ \text{km}^2}\right), Interval(7/10/2014\ 15:25,\ 7/10/2014\ 15:33), Id(2), Id(\text{tornado}))$.

Let S_1 and S_2 be raster granularities such that $S_1 \preccurlyeq S_2$. When a spatial granular term is generalized, a cell or a raster region can remain cell or raster region, correspondingly, but with less precision; or, a raster region can become a cell. The generalization of spatial granular terms is formalized as follows.

A granular term $a_1 = Cell(c)$ of S_1 can be generalized to a granular term $a_2 = Cell(c')$ of S_2 through

$$\mathbb{G}_{\text{Cell}}: (a_1, S_1) \longrightarrow (a_2, S_2) \text{ if and only if } \exists!\ c' \in S_2:\ E(c) \subseteq E(c')$$

that is if there is exactly one granule c' belonging to S_2 such that the extent of c is contained by the extent of c'.

Let $a_1 = RasterRegion(c_1, \ldots, c_n)$ be a granular term of S_1. It can be generalized to a granular term $a_2 = RasterRegion(c_1', \ldots, c_m')$ of S_2 ($m \leq n$) through

$$\mathbb{G}_{\text{RasterRegion}}: (a_1, S_1) \longrightarrow (a_2, S_2) \text{ if and only if } \forall\ i \in \{1\ldots,n\}\ c_i \in S_1\ \exists!j \in \{1\ldots,m\}\ c_j' \in S_2: E(c_i) \subseteq E(c_j')$$

that is if for any granule c_i defining the raster region a_1 there is exactly one granule c_j' belonging to S_2 such that the extent of c_i is contained by the extent of c_j'. Moreover, **the granular term $a_1 = RasterRegion(c_1, \ldots, c_n)$ of S_1 can be generalized to a granular term $a_2 = Cell(c')$ of S_2** through

$$\mathbb{G}_{\text{RasterRegion}}: (a_1, S_1) \longrightarrow (a_2, S_2) \text{ if and only if } \exists!\ c' \in S_2 \forall i \in \{1\ldots,n\}\ c_i \in S_1: E(c_i) \subseteq E(c')$$

that is there is exactly one granule c' belonging to S_2 such that any granule c_i defining the raster region a_1 has its extent contained by the extent of c'.

To illustrate the generalization rules associated with the function symbol *RasterRegion* $\mathbb{G}_{\text{RasterRegion}}$, Figure 4.10 shows two raster regions being generalized to a coarser granularity. The region a_1 changes from a *RasterRegion* to the *Cell* a_2 while the region b_1 remains a *RasterRegion* with less precision denoted as b_2. A study about the generalization of spatial granular terms and its impact on the topological relations (e.g., disjoint, meets, contains) between them was left for future work.

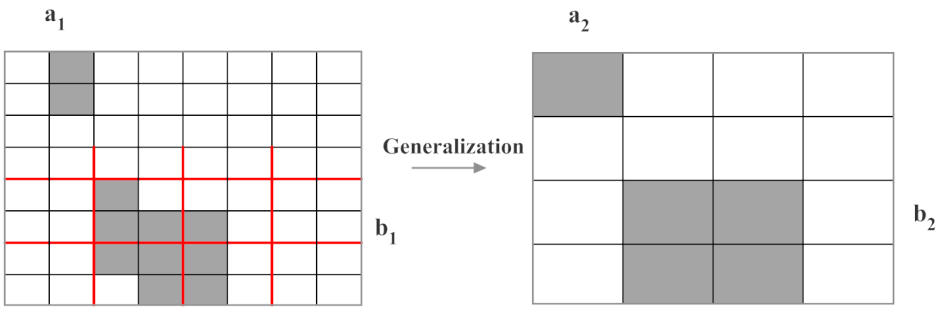


Figure 4.10: Illustration of the generalization rules associated to $\mathbb{G}_{RasterRegion}$.

4.4 Granularities-based Model in Action

The granularities-based model is illustrated with tornadoes occurred in the USA between 1990 and 2015. This phenomenon is described by a collection of 32570 geo-referenced spatiotemporal events. The F1 tornadoes were excluded since their impact in terms of victims is not significant and their spatial coordinates were not accurate, in general. So we kept 27182 spatiotemporal events representing tornadoes with categories ranging from F2 to F5.

These events were modeled through a tornado predicate, with three arguments $tornado(space, time, victims)$. The most detailed spatial granularity $Raster\left(0.13km^2\right)$ is based on a grid of 32768 x 32768 cells that cover the analyzed spatial extent of the phenomenon, and each cell has an area of 0.13 km^2. The other coarser spatial granularities were obtained dividing by a factor of 2 the number of cells in the grid. So the valid granularities for space were rasters with cell sizes of 0.13 km^2, 0.5 km^2, 2 km^2, 8 km^2, 32 km^2. The used time granularities were $Minute$, $Hour$, Day, $Week$, $Month$.

The considered granular terms required to model these events were: Instant and Interval for the time argument; Cell and Raster Region for the space argument.

The raw data (tornadoes) were encoded at the base LoD of the tornado predicate which includes the time granularity of $Minute$ and the space granularity $Raster\left(0.13km^2\right)$. The temporal granular terms Instant and Interval, and the spatial granular terms Cell and Raster Region were used with the tornado predicate according to the data.

As shown in Table 4.1, at the base LoD, some tornadoes were described using:

- the Cell and Instant granular terms (27%) - those with a very short duration and very little spatial expression;

Table 4.1: Percentage of atoms using the proposed granular terms.

	Instant	Interval	Total
Cell	27%	16%	43%
Raster Region	3%	54%	57%
Total	30%	70%	

- the Cell and Interval granular terms (16%) - those with a very little spatial expression but with a time duration larger than a single minute; the average size of the intervals is 8 minutes and 22 seconds;

- the Raster Regions and Instant granular terms (3%) - the few ones that have a duration not larger than a minute with a spatial expression that requires more than one Cell; the average number of cells for the raster regions is 70.6;

- the Raster Regions and Interval granular terms (54%) - the few ones that have a duration larger than a minute and with a spatial expression that requires more than one Cell; the average number of cells for the Raster Regions is 39.6 and the average number of minutes for the Intervals is 7 minutes and 12 seconds.

Notice that, most of the tornadoes (70%) require a granular term Interval. Also, most tornadoes (57%) require a granular term Raster Regions. The description of those tornadoes would be impossible, or at least very hard to encode without the concept of granular terms and especially the Intervals and Raster Regions.

The generalization rules presented in Section 4.3.1 and 4.3.2 were adopted, enabling the automatic computation of the model at coarser LoDs. Given all tornadoes encoded at the base LoD of the tornado predicate with the appropriate granular terms, the model has been computed at coarser LoDs, at all combinations of space and time granularities.

To illustrate how the granularity affects our perception about temporal topological relationships between atoms, let's consider the three F4 tornadoes that occurred in western Iowa on May 27, 1995. The first one (identified as A) started at 18:22 and ended at 19:47. The second one (identified as B) occurred from 18:55 until 20:24. The third one (identified as C) has started at 18:56 and ended at 20:08. At granularity Minutes, A overlaps B and A overlaps C. However, when the tornados are observed at granularity Hours, our perception is changes and therefore A starts B and A starts C.

To study the co-occurrence of tornadoes in space, we compute the total atoms that exist in each spatial granule considering LoDs where atoms (the space argument) are defined using the granularities $Raster\left(0.5km^2\right)$, $Raster\left(8km^2\right)$ and $Raster\left(32km^2\right)$.

For each scenario, we display on a map (Figure 4.11, Figure 4.12, and Figure 4.13) the spatial granules colored according to the number of atoms in it. Orange shows low values while the pink and dark blue ones show high values. Looking at Figure 4.13, the spatial co-occurrence of tornadoes becomes clear in LoDs where the spatial granular terms are built using granules from $Raster\left(32km^2\right)$. Notice that, the change in perception is the result of the change of the granularities and not a result of a change of the classes used in the thematic maps.

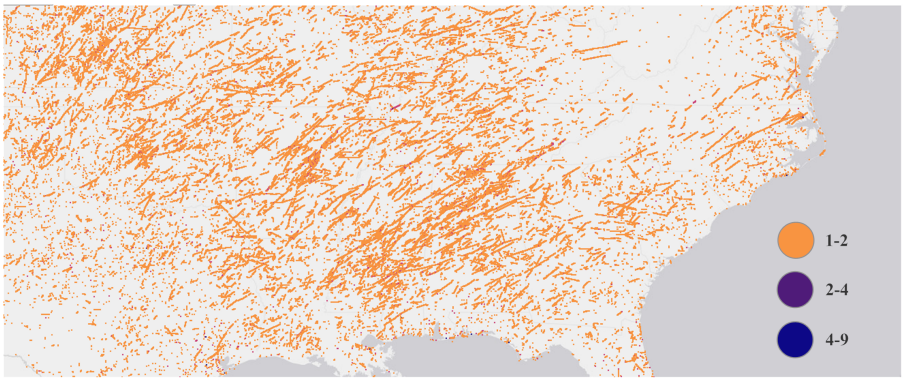


Figure 4.11: An overview of all atoms in LoDs containing the granularity $Raster\left(0.5km^2\right)$.

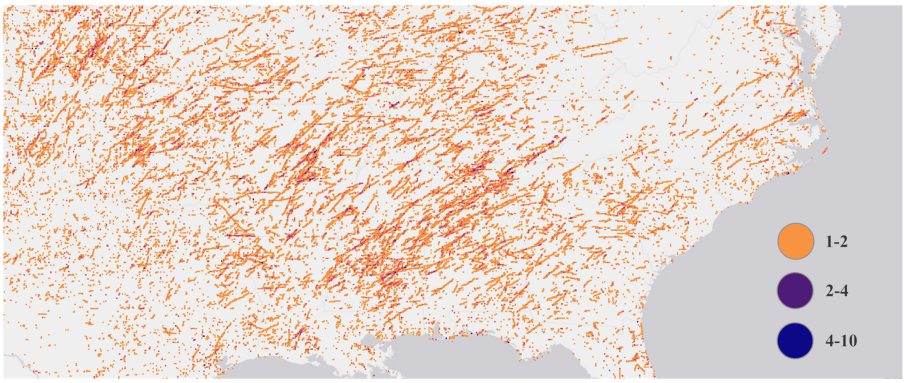


Figure 4.12: An overview of all atoms in LoDs containing the granularity $Raster\left(8km^2\right)$.

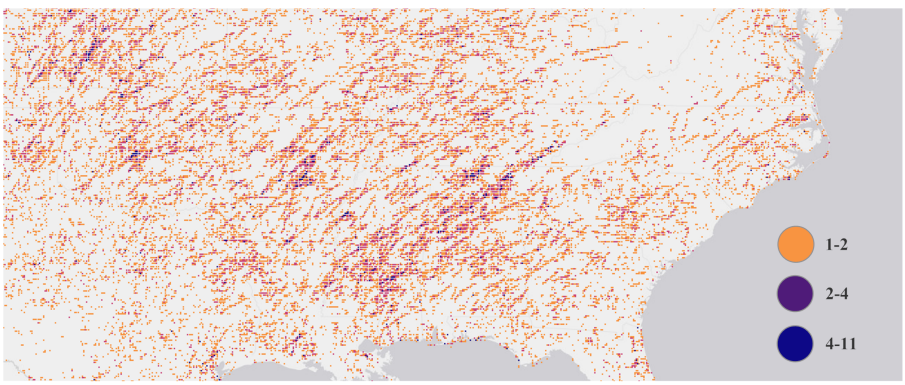


Figure 4.13: An overview of all atoms in LoDs containing the granularity $Raster\left(32km^2\right)$.

4.5 Related Works and their Limitations

This Chapter presents the granularities-based model in order to provide different phenomenon representations for different LoDs. To the best of our knowledge, the concept of LoD (sometimes mentioned in the literature as scale) has been used without being associated to a formal meaning. A core contribution of this PhD Thesis is the formal concept of Level of Detail (LoD) which is the foundation of the granularities-based model on providing different phenomenon representations for different LoDs.

Also, the granularities-based model stands out from others because: (*i*) each predicate provides a representation of the phenomenon like the multirepresentation approaches, but unlike them, there is no need to define everything at the instances level (see Section 2.3.1); (*ii*) unlike the multiresolution approaches, the granularities-based model can express a phenomenon in several LoDs, and not just in several spatial LoDs (see Section 2.3.2); (*iii*) as opposed to current granular computing approaches which are mainly concerned with indexing and aggregating data at different granularities, the granularities-based model provides different phenomena representations for each LoD; also, we provide instruments to create granular syntheses and not just a way of converting information from one granularity to another (see Section 2.3.3); finally (*iv*) once the atoms at the lowest LoD of the predicates are produced, the phenomenon can be expressed into other coarser LoDs automatically based on the atom generalization that relies on the generalization rules for each function symbol; (*v*) based on the general concept of granular term, spatial granular terms (Cell and RasterRegion) and temporal granular terms (Instant and Interval) were formalized; these embedded the generalization-reduction process which is commonly discussed concerning the generalization of spatial features and, consequently, map generalization (see Section 2.3.2); (*vi*) last but not least, the granularities-based model can be easily extended in order to model other kinds of data. To the best of our knowledge, there is no other model combining such characteristics.

Chapter 5

SUITE: A framework for SUmmarizIng spatioTemporal Events

Spatiotemporal events represent a spatiotemporal dynamism that may follow a pattern or embody several patterns. These can be seen like non-identical distributions of events that happen across the entire space and overall time. Finding such patterns can explain or at least can help to understand the phenomena, which can be important for several organizations.

When there is little information about a spatiotemporal phenomenon a user will likely face difficulties during the analysis of phenomena logged as spatiotemporal events. The VA analytical tools are commonly designed to look for non-spatiotemporal patterns based on a single LoD analysis approach so that the difficult choice of the LoDs is left for the users (see Section 2).

However, there are numerous spatiotemporal events collected at high LoDs and the highly dynamic environment embedded in spatiotemporal events provides opportunities to get spatiotemporal patterns in many different forms, perceptible in some LoDs but undetectable in others. Therefore, from our perspective, to enhance the analysis of spatiotemporal events, a user should be provided with an overview about the potential spatiotemporal patterns and the suitable LoDs to find them at early stages of the analysis.

To meet this need, we have proposed a Theory of Granularities (ToG) which is the foundation of the Granularities-based Model. Using the granularities-based model, we can have a phenomenon's representation for each LoD. As spatiotemporal events are being collected at high LoDs, there are many LoDs from which one can analyze the data. However, at this point, we cannot provide an understandable high-level overview about potential patterns across LoDs.

This Chapter proposes a framework for **SU**mmariz**I**ng spatio**T**emporal **E**vents (SUITE) to help users to explore phenomena logged as spatiotemporal events across multiple LoDs,

simultaneously, helping them to understand in what LoDs patterns may emerge. SUITE is devised to build summaries, at different LoDs. The users should be able to inspect and compare the phenomenon's perception across multiple LoDs.

The proposed framework is based on the granularities-based model but, since we are assuming spatiotemporal events, we consider that each used predicate has one and only one argument *space* describing the spatial location of the event, and one and only one argument $time$ specifying the time moment. Other arguments can be used to detail what has happened.

The signature of *event* follows the pattern, $event((space, (G_{(event,\ space)}, F_{(event,\ space)})), (time, (G_{(event,\ time)}, F_{(event,\ time)})), Args)$, and $Args = \{arg_1, (G_{(event,arg_1)}, F_{(event,arg_1)}), \ldots, (arg_n, (G_{(event,arg_n)}, F_{(event,arg_n)})\}$ represents the signature for the other arguments. We also assume that any valid spatial granularity does not have a temporal evolution, i.e., the spatial granularities used remain stable along the temporal scope considered. Furthermore, for the sake of simplification and assuming atoms of the predicate *event*, we will use the following notation: $(\tau)/fr$ to refer to granular syntheses of the *event* predicate in substitution of the granular synthesis notation $G_{Syn}(P(\tau), fr)$ introduced in Chapter 4.

Let $\gamma = \{(space, S), (time, T), \ldots, (arg_n, G_n)\} \in \mathfrak{L}^{event}$ be a *LoD* of *event*. Let $s \in S$ be a spatial granule of the granularity S and $t \in T$ be a temporal granule of the granularity T. Therefore, we can have the pair (s, t) called **spatiotemporal granule** of the spatial and temporal granularity S and T, correspondingly.

Given a spatial and temporal granularity, we may represent the set of spatiotemporal granules based on a simplified representation, like a cube, as displayed in Figure 5.1 where the spatial granularity is illustrated using X and Y axes and the temporal granularity is depicted using the Z axis. The entire extent of a specific cube's cell illustrates a spatiotemporal granule.

A well-formed atom $event\big((space, f_\alpha), (time, f_\beta), args\big)/fr$ represents, at γ LoD, *fr* spatiotemporal events that occurred on spatiotemporal granules referred by a spatial granular term made through a function symbol f_α and a temporal granular term built based on a function symbol f_β such that $f_\alpha \in F_{(event,\ space))}$ and $f_\beta \in F_{(event,\ time))}$. Given a well-formed atom at γ LoD:

- the argument space will be a spatial granular term that might refer to one or more spatial granules
- the argument time will be a temporal granular term that might refer to one or more temporal granules
- the *fr* value might be greater or equal to one

In Figure 5.2, several graphic representations of atoms in terms of their spatiotemporal granules are given.

Figure 5.2a shows an atom with only one spatial granule and one temporal granule, and therefore, one spatiotemporal granule. Figure 5.2b displays an atom with two spatial

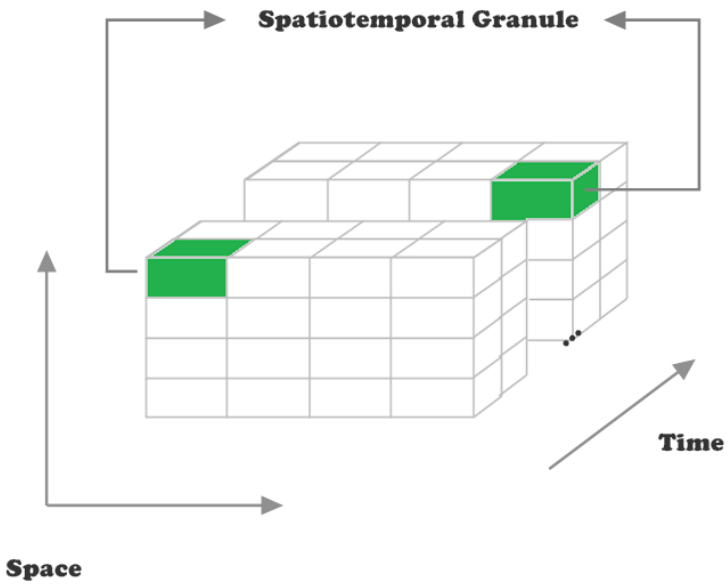


Figure 5.1: Schematic representation of spatiotemporal granules.

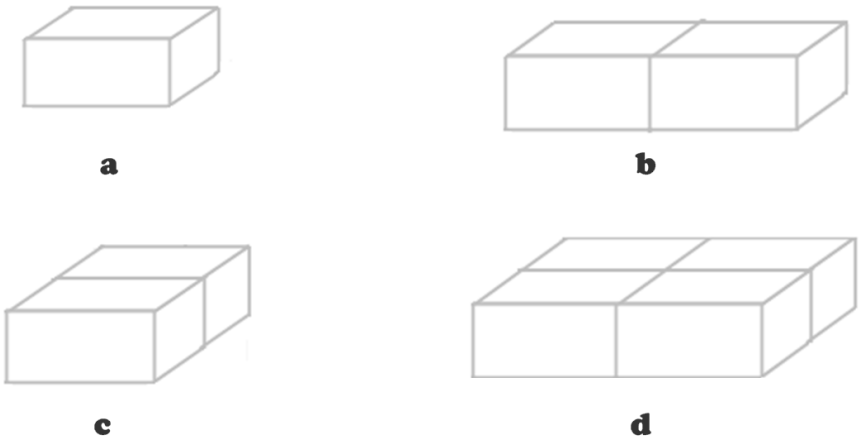


Figure 5.2: Several graphical representations of atoms in terms of their spatiotemporal granules.

granules and one temporal granule that leads to the occupation of two spatiotemporal granules. In Figure 5.2c an atom also occupies two spatiotemporal granules but in this case, it refers to one spatial granule and two temporal granules. Finally, Figure 5.2d displays an atom referring to two spatial granules and two temporal granules and therefore occupying four spatiotemporal granules.

Moreover, when $fr = 1$ is an atom representing only one event. Otherwise, there are fr events that refer to the same spatiotemporal granules and have the exact same description in terms of the other attributes $Args$.

A granularities-based model (or just model), $\mathcal{M}(event)^{\gamma}$, regarding a predicate $event$ at γ LoD can be described as a set of indexed collections of atoms, each indexed by a spatiotemporal granule from $S \times T$ -

$$\Big\{st \rightarrow \Big\{event((space, f_{\alpha}(...,s,...)), \Big(time, f_{\beta}(...,t,...)\Big), args)/fr\Big\} \mid st = (s,t) \in S \times T\Big\} \quad (5.1)$$

In general, a set of atoms is associated to each spatiotemporal granule $st \in S \times T$. This set may be empty, meaning that no event happened at the spatiotemporal granule st; or the set has just one atom, meaning that fr similar events happened at the spatiotemporal granule st; or the set has many atoms, meaning that many different events happened at the spatiotemporal granule st.

The interpretation of what the spatiotemporal granules are indexing might change according to the phenomenon. In case the phenomenon is described only by atoms that occur in only one spatiotemporal granule (see Figure 5.2a) then we can say that spatiotemporal granules are indexing the events that occurred totally in it. When the phenomenon contains atoms described by one or more spatiotemporal granules then we can only say that the spatiotemporal granules are indexing the events that occurred partially in it.

Let's represent by s_{ij} the spatial granules from S_1, t_k the temporal granules from T_1 and $G_{arg} = \{g_a, g_b, g_c\}$. Figure 5.3 presents a set of available atoms indexed by each spatiotemporal granule $(s_{ij},\ t_k)$. For the sake of simplification, in the argument $space$ are only used granular terms of type $Cell$, in the argument $time$ granular terms of type $Instant$ and in the argument agr granular terms of type Id. The functions symbols are omitted in the following formulas.

Each atom is written in a simplified form, such that $event\Big((space, s_{ij}), (time, t_k), (arg, g_{arg})\Big)/fr$ is just represented by g_{arg}/f. For instance, the set of atoms associated with $(s_{13},\ t_1)$ is $\{g_a/4\}$ and $\{g_b/2\}$, and the set of atoms associated with $(s_{24},\ t_1)$ is $\{g_a/1, g_b/3, g_c/1\}$.

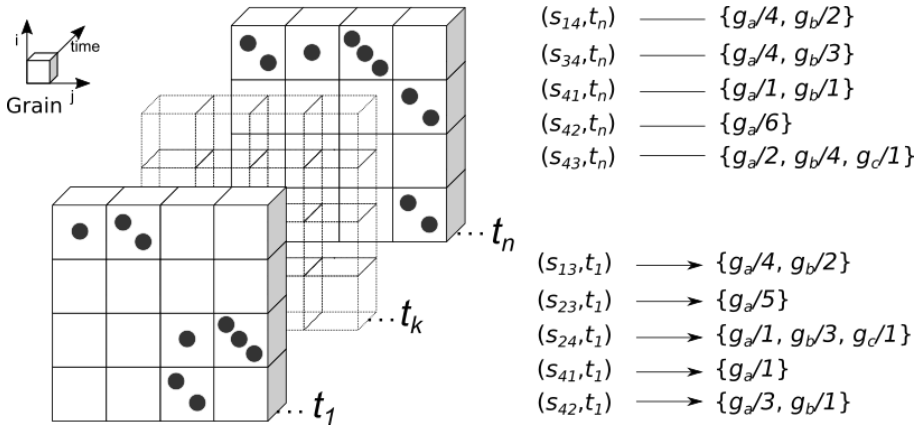


Figure 5.3: Schematic representation of $\mathcal{M}(event)^{\gamma}$.

5.1 SUITE's Overview

The USA traffic accident dataset[1] can be modeled with the granularities-based model using the following predicate $accident(space, time, victims)$. Let's consider the spatial granularities $Raster(0.14\ km^2)$, $Raster(2.27\ km^2)$, $Raster(36.39\ km^2)$, $Counties$, $States$ for the argument $space$; the temporal granularities Day, $Week$, $Month$, and $Year$ for the argument $time$; and, the granularity $Natural\ Numbers$, defined over $\mathbb{N}$ where each granule corresponds to an element of the corresponding domain, for the argument $victims$. For this dataset, we just need the identity function symbol for all the arguments of the accident predicate. The raw data (accidents) were encoded at the base LoD of the accident predicate. Afterward, the generalization of all accidents was done automatically for all LoDs following the needed generalization rule.

This way, the traffic accident dataset can be described at each LoD γ by an equation similar to 5.1, i.e., $\mathcal{M}(accident)^{\gamma}$, where each spatiotemporal granule $st = (s,\ t)$ indexes a set of atoms representing the accidents which happened at that spatiotemporal granule. We can apply simple statistics to summarize $\mathcal{M}(accident)^{\gamma}$.

For instance, some spatiotemporal granules st index empty sets while others index non-empty sets. The percentage of spatiotemporal granules with non-empty sets, named *occupation rate*, measures the average density of a model at a given LoD. Figure 5.4 shows the occupation rate for different spatiotemporal LoDs. On the left-hand chart, the occupation rate is shown for all the spatiotemporal LoDs where the spatial granularity is a raster one. On the chart on the right, the occupation rate is shown for all the spatiotemporal LoDs where the spatial granularity represents an administrative division like $Counties$.

As we can see in Figure 5.4, the occupation rate increases with coarser spatiotemporal granules as expected. But considering just the chart on the left, the spatiotemporal LoD $(Raster(36.39\ km^2), Years)$ has the occupation rate value much greater than the others spatiotemporal LoDs.

[1] USA car accidents occurred between 2001 and 2013, which corresponds to about 450.000 georeferenced accidents: http://www.nhtsa.gov/FARS

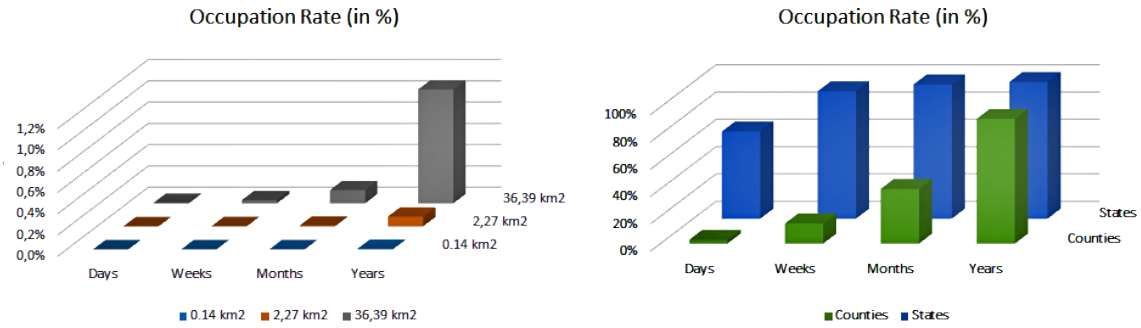


Figure 5.4: The occupation rate for different combinations of spatial and temporal granularities.

At each LoD, the context for the occupation rate, shown in Figure 5.4, is global in the sense that it considers all spatiotemporal granules. The same computation can be done for each temporal granule t_i, considering all the spatiotemporal granules $st = (s, t_i)$. In that case, we get the temporal evolution for the occupation rate computed at each spatial context. This was employed in our data regarding traffic accidents in the USA which result in a time-series for each spatiotemporal LoD as displayed in Figure 5.5. The time series are ordered by the spatial granularity and then by the temporal granularity. The ordering is ascending and is based on the average extent of the granularity's granules.

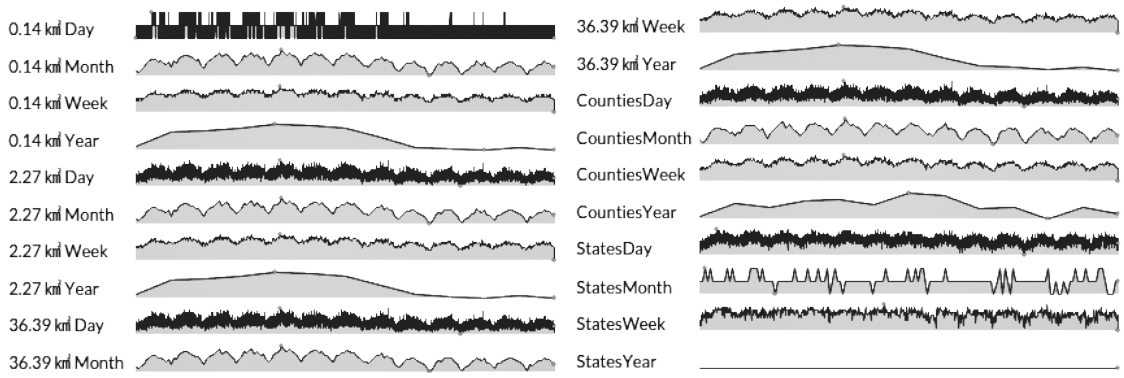


Figure 5.5: The occupation rate computed at each temporal granule.

Each time series is displayed based on its maximum and minimum values. Bearing this in mind, at *States Year* spatiotemporal LoD the one value is a constant which means that there is at least one accident in each state for each year. Another pattern can be seen at 36.39 *Year* spatiotemporal LoD, for instance, which is showing a decreasing trend, meaning that the number of spatial granules with occurrences of accidents has decreased; and, at the spatiotemporal LoDs, containing the granularities *Months* like for example the spatiotemporal $0.14km^2$ *Month* a cyclical pattern is observed that occurs every year. At the end of February, the number of traffic accidents reach, in general, its minimum value and then starts to increase. Around September the number of traffic accidents starts to decrease.

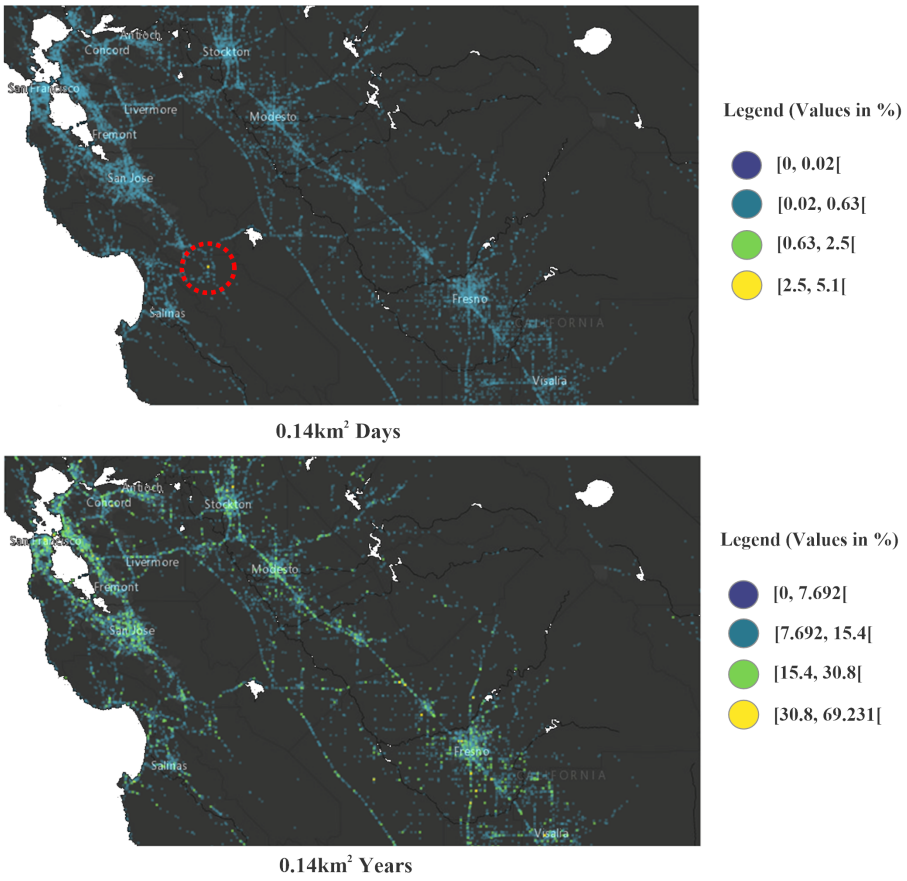


Figure 5.6: The occupation rate computed at each spatial granule.

On the other hand, the occupation rate computation can also be done for each spatial granule s_j, considering all the spatiotemporal granules $st = (s_j, t)$, getting for each spatial granule the occupation rate across all the temporal granules. Figure 5.6 shows two maps where each "point" represents a spatial granule and its color is given by the occupation rate value according to the map's legend (see Figure 5.6).

The map at $0.14km2$ *Days* spatiotemporal LoD shows an outlier, highlighted by a dashed circle. In the "yellow" spatial granule, there are accidents occurring with some degree of frequency in comparison with the other granules. When we change the spatiotemporal LoD to $0.14km2$ *Years*, the perception is changed and that outlier is no longer perceived.

The proposed framework builds summaries of each phenomenon's LoD to support users in carrying inspection and comparison tasks of a phenomenon across multiple LoDs. Observing summaries across multiple LoDs can provide useful information to identify the proper ones to carry out a particular analysis. Generically, we will refer to those summaries as *abstracts*.

5.2 Abstracts

Our framework was designed to build abstracts over $\mathcal{M}(P)^{\gamma}$. An abstract $\mathbb{A}$ can be, for instance, a number, a vector, or even a matrix measuring a particular feature of a phenomenon. Five types of abstracts with different contexts are introduced: (*i*) Global Abstract; (*ii*) Temporal Abstract; (*iii*) Spatial Abstract; (*iv*) Compacted Temporal Abstract; (*v*) Compacted Spatial Abstract.

A Global Abstract is a single summary of all atoms indexed by spatiotemporal granules as illustrated in Figure 5.7. For example, in Figure 5.4, we have displayed one Global Abstract (i.e., the occupation rate) for each spatiotemporal LoD about traffic accidents in USA. The Global Abstract is formally defined as follows.

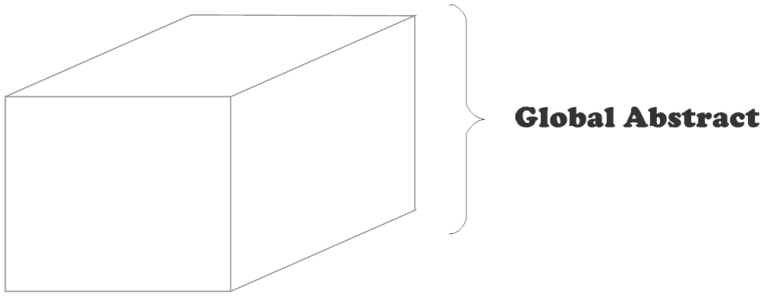


Figure 5.7: The intuition of the Global Abstract.

Definition 5.1 (Global Abstract)**.** Let $\mathcal{M}(P)^{\alpha}$ be the set of granular syntheses indexed by each spatiotemporal granule. Thus, a function $\mathbb{F}_{Global}{:}(\mathcal{M}(P)^{\alpha}) \rightarrow \mathbb{A}_{Global}$ produces a global Abstract such that $\mathbb{A}_{Global}$ is one abstract $\mathbb{A}$.

Spatiotemporal statistics can be used to produce Global Abstracts (see Section 2.1.3) in order to get hints about properties concerning the distribution of spatiotemporal events.

Global Abstracts may hide some important variations in space and/or time. Hence, we introduce the possibility to create abstracts that are more "detailed". One of them is the Spatial Abstract.

A Spatial Abstract contains a summary for each temporal granule. The intuition of this type of abstract is given in Figure 5.8.

As an example several Spatial Abstracts were shown in Figure 5.5 (i.e., the occupation rate), one for each spatiotemporal LoD. The Spatial Abstract is formally defined as follows.

Definition 5.2 (Spatial Abstract)**.** Let $\mathcal{M}(P)^{\gamma}$ be the set of granular syntheses indexed by each spatiotemporal granule. Thus, a function $\mathbb{F}_{Spatial}{:}(\mathcal{M}(P)^{\gamma}) \rightarrow \mathbb{A}_{Spatial}$ produces an abstract for each temporal granule such that $\mathbb{A}_{Spatial} = \{(t, \mathbb{A}) \mid t \in \mathrm{T}\}$.

A Spatial Abstract is a summary based on $\mathcal{M}(P)^{\gamma}$ for each $t \in T$. It allows us to look at the evolution of a summary over time, which is measuring a **spatial feature** of a

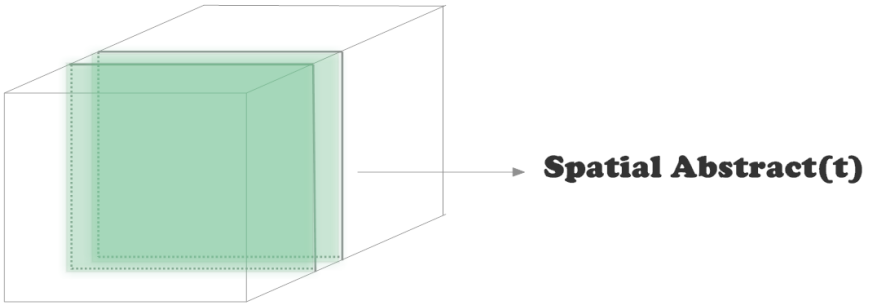


Figure 5.8: The intuition of the Spatial Abstract.

phenomenon. For example, spatial statistics can be used in order to understand the spatial distribution of events in each temporal granule. This way, a user can assess whether the events occurred in a dispersed form or if they happened in a clustered manner. This might be particularly useful to capture the temporal non-stationary of spatiotemporal events.

As the Spatial Abstract allows one to look at a summary over time, we introduce the Temporal Abstract to look at a summary over space. In this case, a summary for each spatial granule is computed as illustrated in Figure 5.9. Two examples of Temporal Abstracts were provided in Figure 5.6, in which the occupation rate was computed. The Temporal Abstract is formally defined as follows.

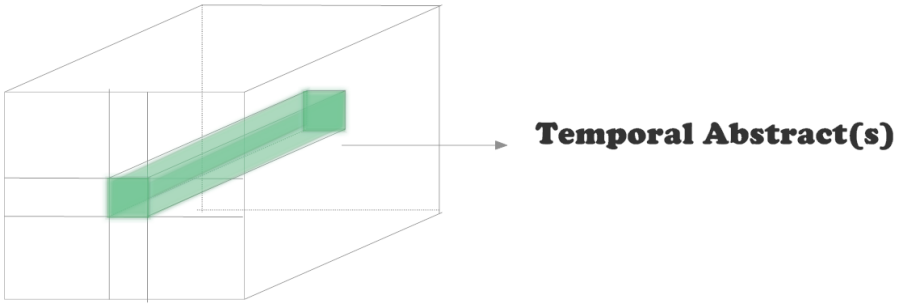


Figure 5.9: The intuition of the Temporal Abstract.

Definition 5.3 (Temporal Abstract)**.** Let $\mathscr{M}(P)^{\gamma}$ be the set of granular syntheses indexed by each spatiotemporal granule. Thus, a function $\mathbb{F}_{Temporal}{:}(\mathscr{M}(P)^{\gamma}) \rightarrow \mathbb{A}_{Temporal}$ produces an abstract for each spatial granule such that $\mathbb{A}_{Temporal} = \{(s, \mathbb{A}) \mid s{\in}\mathrm{S}\}$.

A Temporal Abstract is a summary based on $\mathscr{M}(P)^{\gamma}$ for each $s \in S$. It allows us to look at a summary over space, which is measuring a **temporal feature** of a phenomenon. For example, temporal statistics can be used so that we are able to understand the temporal distribution of events in each spatial granule. This way, for each spatial granule, one can assess if events occurring on a particular spatial granule are close or dispersed to each other in time.

Moreover, each Spatial (or Temporal) Abstract can be further summarized into a single summary that we called Compacted Spatial (or Temporal) Abstract.

Definition 5.4 (Compacted Spatial Abstract)**.** Let $\mathbb{A}_{Spatial}$ be a Spatial Abstract. Thus, a function $\mathbb{F}_{CompactSpatial}{:}\left(\mathbb{A}_{Spatial}\right) \rightarrow \mathbb{A}_{CompactSpatial}$ produces a Compacted Spatial Abstract such that $\mathbb{A}_{CompactSpatial}$ is one abstract $\mathbb{A}$.

For each Spatial Abstract (i.e., time series) displayed in Figure 5.5, we can use an aggregation measure, like the average, to produce a Compacted Spatial Abstract. Other methods that come from descriptive statistics or methods to analyze time series can be used to build Compacted Spatial Abstracts.

Definition 5.5 (Compacted Temporal Abstract)**.** Let $\mathbb{A}_{Temporal}$ be a Temporal Abstract. Thus, a function $\mathbb{F}_{CompactTemporal}{:}\left(\mathbb{A}_{Temporal}\right) \rightarrow \mathbb{A}_{CompactTemporal}$ produces a Compacted Spatial Abstract such that $\mathbb{A}_{CompactTemporal}$ is one abstract $\mathbb{A}$.

For each Temporal Abstract (i.e., map) displayed in Figure 5.6, we can also use an aggregation like the average to produce a Compacted Temporal Abstract. Other methods that come from descriptive statistics or spatial statistics can be used to produce Compacted Temporal Abstracts.

To wrap up all the types of abstracts proposed, there is an example displayed in Figure 5.10. This example assumes a spatial granularity with 16 spatial granules (4x4) and a temporal granularity with four temporal granules (i.e., a particular spatiotemporal LoD). Thus, we have 16x4 spatiotemporal granules that are marked with one when they were occupied by some atom.

Within the "red" area, the occupation rate is displayed as **global abstract** which consists of the value 17.2. The occupation rate as **spatial abstract** is displayed within the "purple" area as well as its average (i.e., **compact spatial abstract**). Finally, the occupation rate as **temporal abstract** is displayed within the "blue" area as well as its average (i.e., **compact temporal abstract**).

5.3 Properties of Abstracts Functions

Abstracts are built through functions. Each function will measure one feature of the phenomenon which in turn can employ different strategies using different information from the model $\mathcal{M}(P)^{\gamma}$. Bearing this in mind, we identified three properties that can further characterize the function that computes an abstract. These properties describe the way each spatiotemporal granule contributes to the Abstract computation, i.e., the way each $\eta = st \rightarrow \left\{\left(((space, f_{\alpha}(...,s,...)), \left(time, f_{\beta}(...,t,...)\right), args\right)/fr\right\}$ is integrated for the resulting abstract. They are: (*i*) neighborhood dependency; (*ii*) spatiotemporal dependency; (*iii*) semantic dependency. These properties are further detailed.

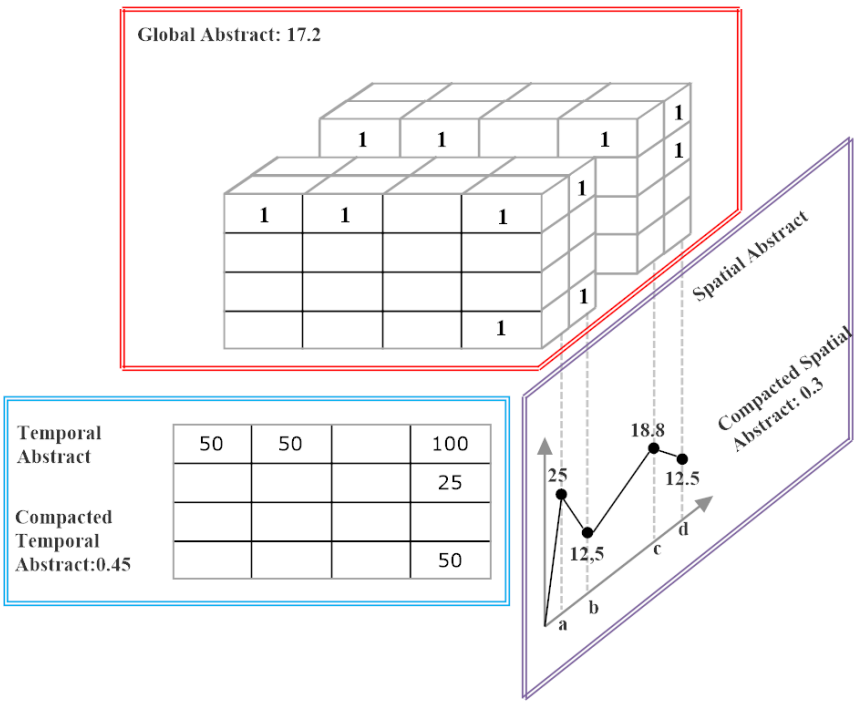


Figure 5.10: Summary of all Types of Abstracts.

Neighborhood dependency. The contribution of each η for the *Abstract* depends (or not) on the spatiotemporal neighborhood. This neighborhood dependency may be only temporal (e.g., depends only on the events that happen on their neighbor temporal granules); only spatial (e.g., depends only on the events that happen on their neighbor spatial granules); or may be both spatial and temporal-dependent. For instance, popular methods that measure spatiotemporal interaction like Knox and Bartlett 1964, Mantel 1967, Jacquez 1996 k Nearest Neighbor can be used as global abstracts that are spatial and temporal-dependent.

When an *Abstract* computation is not dependent then the computation of $\mathbb{F}(\{\eta\})$, where $\eta = st \rightarrow \left\{((space, f_\alpha(...,s,...)), \left(time, f_\beta(...,t,...)\right), args)/fr\right\}$, can be rewritten as $\mathbb{F}(\{\eta\}) = Agg_{\mathbb{F}}(\{\mathbb{F}'(\eta)\})$, where $\mathbb{F}'$ computes the contribution of each η and $Agg_{\mathbb{F}}$ aggregates those contributions to get the final *Abstract*. This means that $\mathbb{F}'$ can be a function of local computation not requiring information about others η.

Spatiotemporal dependency: the contribution of each η for the *Abstract* depends (or not) on the specific spatiotemporal granules $st = (s,t)$ of η. This spatiotemporal dependency may be only temporal (e.g. the contribution is different if the events happened at night or during the day, or even varying with the season); only spatial (e.g., the contribution is different if the events happened at high mountains or at sea level, or even varying according to the spatial granule like the specific counties); or may be both spatial and temporal dependent.

When an *Abstract* computation is not spatiotemporal dependent then the computation of $\mathbb{F}(\{\eta\})$, where $\eta = st \rightarrow \left\{((space, f_\alpha(...,s,...)), \left(time, f_\beta(...,t,...)\right), args)/fr\right\}$, can be rewritten. Consider η' as $\eta' = st \rightarrow \{event(args)/fr\}$ where we removed the information about *space* and $time$ and leave the set $\{event(args)/fr\}$ indexed by st just to keep any neighborhood information between spatiotemporal granules. Then, $\mathbb{F}(\{\eta\}) = \mathbb{F}(\{\eta'\})$.

When an *Abstract* computation is neither spatiotemporal dependent nor neighborhood dependent, $\mathbb{F}(\{\eta\})$, it can be rewritten as $\mathbb{F}(\{\eta\}) = Agg_{\mathbb{F}}(\{\mathbb{F}'(\{\eta'\})\})$, where $\mathbb{F}'$ computes the contribution of each set $\{\eta'\}$ independently of their spatiotemporal location and $Agg_{\mathbb{F}}$ aggregates those contributions to get the final *Abstract*.

Semantic dependency: the contribution of each η for the *Abstract* depends (or not) on the semantic arguments of η. When the *Abstract* is not semantic dependent then $\eta = st \rightarrow \left\{((space, f_\alpha(...,s,...)), \left(time, f_\beta(...,t,...)\right), args)/f\right\}$ can be simplified to $\eta = st \rightarrow \{event((space, f_\alpha(...,s,...)), (time, \beta(...,t,...)))/fr\}$. For instance, if we are studying car accidents, an *Abstract* semantic dependent will consider the type of accident and/or the number of victims, while an *Abstract* semantic independent only considers the number of accidents.

When an *Abstract* computation is neither spatiotemporal dependent nor neighborhood dependent nor semantic- dependent, $\mathbb{F}(\{\eta\})$, it can be rewritten as $\mathbb{F}(\{\eta\}) = Agg_{\mathbb{F}}(\{\mathbb{F}'(\{fr\})\})$, where $\mathbb{F}'$ computes the contribution of each bag $\{fr\}$, and $Agg_{\mathbb{F}}$ aggregate those contributions to get the final *Abstract*. The occupation rate is an extreme example of such *Abstract* and can be defined based on: $\mathbb{F}'(\{\eta\}) = if \left|\{\eta\}\right| > 0$ *then* 1 *else* 0:

$$Agg_{\mathbb{F}}(\{\mathrm{x}\}) = \frac{\sum x}{|S \times T|} \tag{5.2}$$

5.4 Discussion

Our framework allows us to define or use many functions available in the literature that create summaries of data.

As presented, the functions computing abstracts may be semantic dependent. Such dependency is delimited by the predicate's signature regarding the arguments *args*. These arguments depend on the phenomenon itself. In the case of car accidents, one may collect information about the number of victims, whether some of the drivers present an alcoholic rate above the legally allowed, information about weather conditions, among others. A function computing an abstract can use this information. For instance, one can compute the occupation rate by weather conditions as Global Abstract; or we can use the Global Moran's I Moran 1950 to build a Spatial Abstract that measures the correlation between spatiotemporal granules and the weather conditions.

Furthermore, the functions producing abstracts may be spatial and/or temporal dependent. In case of dependency, it is important to have a base knowledge for each spatial and temporal granule and that base knowledge should be relevant for the phenomenon

in study. Some examples to describe a temporal granule are: the time of day that each temporal granule occurs (e.g., night or day), what kind of season it is in. The spatial granules can be characterized, for instance, as information about altitude, if is a rural or urban area.

Moreover, the functions computing abstracts may be neighborhood dependent. This dependency can make the functions more time-consuming when compared with the neighborhood independent ones. Some examples are given: (*i*) Gabriel et al. Gabriel et al. 2013 estimators or the average nearest neighbor index Ebdon 1985 can be used as Global Abstracts in order to measure the spatiotemporal clustering/regularity of spatiotemporal granules; (*ii*) the average nearest neighbor index Ebdon 1985 can also be used as Spatial Abstract computing for each temporal granule a clustering measure which might indicate variations between dispersed and clustered spatial distributions; alternatively, it may reveal constant dispersed or clustered distributions; (*iii*) Keogh et al Keogh et al. 2005 propose an algorithm to find the most unusual subsequence within a time series, which can be used as Temporal Abstract. Such abstract is computed by a function temporal neighborhood dependent; (*iv*) based on the Fourier discrete transform, a function may compute a Temporal Abstract returning the n higher frequencies. Such function is temporal neighborhood dependent.

In the absence of the neighborhood dependency, functions making abstracts can work individually for each spatiotemporal granule as discussed. For this reason, parallel computing techniques can be employed.

Finally, the functions computing abstracts may be used in different abstracts holding different properties and sometimes they depend on the phenomena in study. However, it is fundamental that those functions provide comparable abstracts. By comparable abstracts, we mean abstracts that are not completely sensitive to the spatiotemporal LoD at which the abstract is being computed. For example, the occupation rate is an example of an abstract that is influenced by the size of the spatiotemporal granules as it was observed in Figure 5.4. Therefore, little information can be extracted from it as the value of the occupation rate will likely growth as long as coarser spatiotemporal LoDs are considered.

Ultimately, we aim to support users carrying inspection and comparison tasks of a phenomenon across multiple LoDs. To this end, comparable abstracts are advised to allow a fair comparison among phenomenon's LoDs.

5.5 Main Abstracts Implemented

Several abstracts were implemented and actually proposed in the context of this work. Whenever some abstract is based on another work a reference will be placed. By default, i.e., if nothing was said you can assume that it was proposed in this work. A subset of abstracts implemented/proposed are described:

1. The **Occupation rate** measures the percentage of spatiotemporal granules occupied, that is, measures the average density of a model at a given LoD. The value 0 means no spatiotemporal granules are occupied and 100 means that all spatiotemporal granules are occupied.

2. The **Collision rate** measures the percentage of the spatiotemporal granules occupied that index more than one atom, that is, it measures the average co-occurrence of a model at a given LoD. In this case, 0 means no co-occurrence of granular syntheses on spatiotemporal granules and 100 means that any granular synthesis is co-occurring at least with another one.

3. The **Granular Mantel Bounded and Normalized (GMBN)** measures the spatiotemporal interaction among granular syntheses. The purpose of this measure is to have a hint of the presence or absence of spatiotemporal clustering pattern or any other pattern that involves spatiotemporal interaction like the contagious process. The value ranges between 0 and 1 where 0 means no interaction at all among the granular syntheses and 1 means that all the granular syntheses are interacting among each other's. The GMBN receives as input parameters the spatial and temporal distances. These distances are expressed in terms of granular extents with respect to the spatiotemporal LoD in which the GMBN is computed. Both parameters were fixed with the value two. For example, if the GMBN is being computed at the spatiotemporal LoD $(Raster(41.74km^2), Days)$ the spatial distance will be 13 km (i.e., twice the $sqrt(41.74)$) and the temporal distance will be 2 days. This abstract is a contribution to handle some limitations found on popular methods to measure spatiotemporal interaction like Knox and Bartlett 1964, Mantel 1967, Jacquez 1996 k Nearest Neighbor. A more detailed discussion about GMBN is provided in Appendix D.

Spatial Abstracts hold a summary for each temporal granule about the granular syntheses occurred on it. The Spatial Abstracts considered are:

1. The **Spatial occupation rate** is computed in the scope of each temporal granule. The values' interpretation is similar to the one presented considering the global abstract. This way, we can track the temporal evolution of the occupation rate.

2. The **Spatial frequency rate** measures for each temporal granule the percentage of atoms occurred on it given all the atoms of the phenomenon at a given LoD. In other words, corresponds to a frequency distribution normalized by the total number of atoms in the phenomenon at particular LoD. The range of values for this abstract lies between 0 and 1 (in each temporal granule) so that 0 means that no atom occurred on that temporal granule while 1 means all the atoms occurred on that temporal granule. Through this abstract, we aim to understand how the

intensity of the phenomenon spreads out throughout time. This is abstract is not a novel contribution.

3. The **Spatial average nearest neighbor (Spatial ANN)** measures how granular syntheses are dispersed or clustered in each temporal granule. This might indicate variations between dispersed and clustered spatial distributions. The value computed is not a distance but a normalized value such that if the value is less than 1, the spatial pattern might be clustering while if the value is greater than 1, the trend is toward dispersion. Notice that, the z-score[2] of the Spatial ANN is also computed. Very low or very high z-score values suggest some spatial pattern, and therefore, we can reject the complete spatial randomness. This abstract was developed based on Ebdon 1985.

4. The **Spatial scope** measures the percentage of spatial area occupied by the phenomenon in each temporal granule, where the spatial area is a concave region that encloses all the granular syntheses, and the total spatial area corresponds to the extent of the spatial granularity. Through this abstract, we aim to understand if the spatial scope of the phenomenon varies throughout time.

5. The next Spatial Abstract, considers two consecutive temporal granules t_{i-1}, t_i. For each one, a region that encloses all granular syntheses is computed. Then, the centroids of each region are computed, and the value of the Spatial Abstract at t_i consists of the distance between the centroid at t_{i-1} and the centroid at t_i. This is done for all temporal granules apart from t_0 where the Spatial Abstract takes the value 0. We call this Spatial Abstract as the **Spatial Consecutive Distance Between Centers of Mass**. Through this abstract, we aim to understand whether the phenomenon moves in space throughout time. This abstract is a contribution of this work, and to the best of our knowledge, similar statistics were not found in the literature.

6. Another way of understanding how the phenomenon moves in space throughout time is to project the centroid's coordinates of the region that encloses all granular syntheses (in each temporal granule) into a one-dimensional domain. One way of doing such a measure is as follows. Let's consider the minimum bounding box that encloses the extent of the spatial granularity so that the upper left corner corresponds to the 0 value and the bottom-right corner corresponds to the 1 value. The value goes from 0 to 1 as long as we move in space from top to bottom and left to right. Therefore, for each temporal granule, the centroid's coordinates of the region that encloses all granular syntheses are projected between 0 and 1 following the mentioned mapping. We call this Spatial Abstract as the **Spatial Center Mass's Positioning**. This abstract is a contribution of this work, and to the best of our knowledge, similar statistics were not found in the literature.

[2] http://mathworld.wolfram.com/z-Score.html

Temporal Abstracts hold a summary for each spatial granule about the granular syntheses occurred considering all temporal scope. The Temporal Abstracts considered are:

1. The **Temporal occupation rate** is, in this case, computed in the scope of each spatial granule. The values' interpretation is similar to the one presented considering the Occupation rate. This way, we can assess the occupation rate over the space.

2. The **Temporal frequency rate** measures for each spatial granule the percentage of atoms occurred on it given all the atoms of the phenomenon at a given LoD. The range of values for this abstract lies between 0 and 1 (in each spatial granule) so that 0 means that no atom occurred on that spatial granule while 1 means all the atoms occurred on that spatial granule. This way, we can observe the intensity of the phenomenon over the space.

3. A Temporal Abstract **Temporal average nearest neighbor** measures how granular syntheses are dispersed or clustered in time for each spatial granule. The interpretation of values is similar to the one presented in the case of the Spatial average nearest neighbor. Furthermore, the corresponding z-score was also implemented.

4. A Temporal Abstract **Temporal center mass's positioning** was also implemented. Given the temporal granularity underlying the computation of this abstract, 0 means that all the granular syntheses occurred on the "first" temporal granule and 1 means that all the granular syntheses occurred the "last" temporal granule. This measure can provide hints about the relation among spatial granules and the time that the events occurred on it.

Finally, Compact Spatial Abstracts and Compact Temporal Abstracts were also considered. For each Spatial Abstract, the average and the coefficient of variation[3] were implemented as Compact Spatial Abstracts. The same was done for each Temporal Abstract that result in Compact Temporal Abstracts.

Furthermore, a Compact Temporal Abstract measuring the spatial autocorrelation of each temporal abstract was implemented. This was done by adapting the Pearson's Correlation Coefficient into the spatial context, and therefore, the value falls in the range of -1 to +1, where being close to -1 indicates strong spatial negative correlation, +1 means strong spatial positive correlation and 0 indicates no spatial correlation.

5.6 Related Works and their Limitations

This Chapter presents the SUITE framework to support users in carrying the inspection and comparison tasks of a phenomenon across multiple LoDs, without having to look

[3]The coefficient of variation is a measure of spread that describes the amount of variability relative to the mean. Because the coefficient of variation is unitless, you can use it instead of the standard deviation to compare the spread of data sets that have different units or different means.

at raw data, and to handle the spatiotemporal complexity. As our framework does not make any assumption about the phenomenon and the analytical task, it can be widely used to get an overview of the phenomenon under analysis. To the best of our knowledge, there are no approaches that work across several spatial and temporal LoDs, and that are independent of the analytical task and the domain, applicable in the context of spatiotemporal events.

Even so, there are approaches to make analyses over spatiotemporal events. In general, standard practices provide tools and approaches that work on a single LoD driven by the user. Geovisualization, automated and visual analytics approaches fit this description (see Section 2.4). However, the LoD plays a crucial role during the analytical process and, often, there is no exclusive LoD to analyze a phenomenon.

Furthermore, those approaches revealed other issues. Usually, applications make use of geovisualization methods to display raw spatiotemporal events. Such an approach makes the perception of patterns in spatiotemporal events challenging, from the human viewpoint, as the users have to handle the spatiotemporal complexity. On the other hand, the automated approaches focused on a particular pattern, and still, effective visualizations need to be used to communicate externally the pattern identified. From our viewpoint, such approaches should be used when there is evidence of those patterns and not at early stages of analysis when little is known about phenomena. Besides, several spatiotemporal patterns may occur in a phenomenon and some of them may be strongly related. Focusing on a particular one can make us miss other patterns that may be present on the data. Recently, visual analytics approaches have been proposed but they are frequently focused on clustering tasks or employ common statistics that do not handle spatiotemporal unique properties of spatiotemporal events.

Spatial and spatiotemporal statistics are developing quantitative analytical methods that provide hints about possible patterns in spatiotemporal events, which can be easily perceived by the end user. However, there was no framework formally defined to frame their computation at several LoDs. With this work, one can have a high-level overview of phenomena at multiple LoDs, simultaneously, something that was no archived so far.

CHAPTER 6

EXPERIMENTS AND RESULTS

A Visual Analytics tool was developed, implementing the granularities-based model as well as the SUITE framework in order to enhance the exploratory analysis of spatiotemporal events. The tool is called SUITE-VA and its architecture, technologies, and interface are presented in Section 6.1. The experiments and results are reported in Section 6.2.

We established five types of abstracts working with space and time together in order to measure different facets of phenomena logged through spatiotemporal events. These abstracts are anchored on a theoretical framework that frame the computation of abstracts at different LoDs. Phenomena are modeled through the *event* predicate that require the *space* and *time* arguments and each abstract is computed for each spatiotemporal LoD. Since the same abstract can be observed in different spatiotemporal LoDs, one can observe the way abstracts vary according to the LoD, and therefore, in what LoDs facets are better perceived.

A particular facet might reflect a pattern per se, or it may be revealed by the joint interpretation of several abstracts. As the framework proposed does not make any assumptions about the phenomenon and the analytical task, it can be widely used to get an overview about the presence or absence of different patterns across LoDs.

The SUITE-VA allows us to visually inspect the abstracts in order to understand the absence or presence of different kinds of spatiotemporal patterns at multiple LoDs, simultaneously, following a coordinated strategy among the visualizations provided. The SUITE-VA is detailed as follows.

6.1 SUITE-VA Tool

SUITE-VA is a web-based tool and follows a client-server architecture. The server is coded in Java providing a set of RESTful Web services (Spring). It relies on the PostgreSQL as

the Database Management System, and its spatial extension PostGis. The browser-based client is coded in JavaScript, HTML5, and uses WebGL to display efficiently thematic maps.

SUITE-VA is composed by three-modules decoupled from each other: (*i*) Granularities-based module; (*ii*) SUITE module; (*iii*) Interface module. The granularities-based and SUITE modules are placed on the server-side while the interface module is placed on the client-side. Furthermore, the granularities-based and SUITE modules are decoupled from the RESTful Web Service. The application server provides a set of services that are implemented using the interfaces exposed by the granularities-based and the SUITE modules. These will be later used by the interface module. An overview diagram for the SUITE-VA architecture is provided in Figure 6.1.

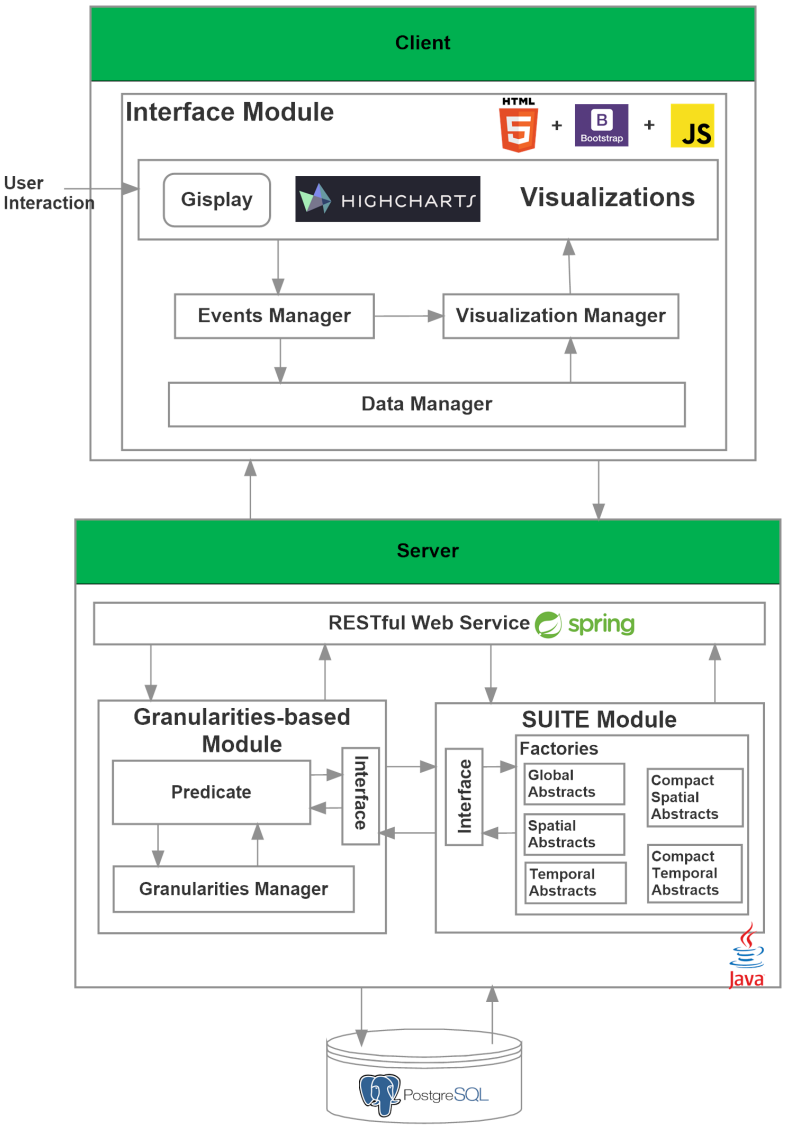


Figure 6.1: The SUITE-VA tool architecture.

The **Granularities-based module** receives as input a dataset of spatiotemporal events

and a predicate signature compliant with the *event* signature presented in Chapter 5, and then it automatically generates the set of atoms for each LoD of the corresponding predicate. The set of LoDs $\mathbb{L}^{event}$ is inferred based on the granularities defined for each argument and the relationship $finer\ than$ that exists between them.

The **Granularities Manager** manages granularities. The information about each granularity is stored in the database like the name, the long name, the short name, the extent, the number of granules as well as the set of granules or the meta-information needed to know the set of granules (i.e., in case of the all granules being regular). Each granularity can be loaded in-memory, and when loaded, it becomes a Java object with a set of functions, namely, one might access to a granule's extent based on its index value, or to check whether the granularity is *finer-than* another.

The **Predicate module** receives information about the arguments as well as the corresponding valid granularities and function symbols. Based on the valid granularities, it produces automatically the set of LoDs available. Afterward, it encodes the spatiotemporal events at the base LoD, and then, the data can be generalized for all the LoDs available. But if a user wants to just generalize the data for a particular LoD in order to export them later on onto a CSV file, it is also possible. Each LoD of the phenomenon is stored on a table of the database. In our implementation, the generalization occurs from the base LoD to a target LoD. Our implementation supports the generalization of the granular terms proposed.

During the generalization's computation, local summaries associated to each spatiotemporal granule are built. For example, the number of spatiotemporal events, the number of distinct atoms, among others. These are used by the abstracts' functions (in the SUITE module) in order to take advantage of the computation already performed when they are executed. Also, database indexes (spatial and nonspatial) are created to improve query execution time. Although the generalization occurs before the user is interacting with data, special attention to the performance was needed, especially during the computation of abstracts in the SUITE module as discussed later.

The tool was tested on datasets up to 1 million events. For datasets with 50.000 events the generalization from the based LoD to a target LoD takes around 30 seconds, which include the generalization of the actual data, the creation of auxiliary tables, and the creation of spatial and non-spatial indexes. For datasets with 1.000.000 events it takes around 30 minutes. Notice that, the times here exposed did not result from an exhaustive evaluation but rather they are just approximate times that provide a grasp about the time the tool takes to perform the task.

The **SUITE module** receives the predicate signature and meta-information about the function symbols implemented to compute abstracts. For each function defined, the computation of all the available LoDs is made. The abstracts computed for each LoD of the *event* predicate are stored persistently. In this module, the local summaries and the indexes created have an important role.

As an example, the Granular Mantel Bounded and Normalized (GMBN) requires, for

each granular synthesis, the computation of its neighbors within a spatial and temporal distance. Thus, let's say that we have a dataset of half a million events and 20 spatiotemporal LoDs. In some LoDs, the computation of the GMBN will require near a half million neighborhoods operations. Performing that operation for all LoDs can be time-consuming and without any indexes, it might take a few days to be completed. For this reason, some extra attention to performance was needed in spite of the abstracts being precomputed. In our implementation, in datasets containing around 30.000 events, the computation of GMBN for a single LoD takes less than 20 seconds while in datasets of half a million events it takes about less than 30 minutes.

The **SUITE module** follows the abstract factory design pattern, and therefore, there is a factory for each type of abstract proposed. This way, new abstracts' functions can easily be added, since the SUITE-VA was developed in a modular way. In the end, one just needs to define a new Java Class. The functions developed to generate abstracts are listed in Appendix C. In short, we implemented 5 Global Abstracts, 16 Spatial Abstracts, 6 Temporal Abstracts. From the 16 Spatial Abstracts result 32 Compact Spatial Abstracts (Average and Coefficient of Variation) and from the 6 Temporal Abstracts result 18 Compact Temporal Abstracts (Average, Coefficient of Variation and Spatial autocorrelation).

The Visual Analytics **Interface module** receives the predicate signature. This information is used by the **Data Manager** to retrieve the set of abstracts precomputed for each spatiotemporal LoD. The **Visualization Manager** receives the abstracts' values for each spatiotemporal LoD that are turned into visualizations. This module keeps track of the visualizations being displayed.

As we follow a coordinated strategy among the visualizations provided, the **Event Manager** handles the events triggered by the user interactions on the visualizations displayed. Then, the actions needed to keep the visualizations coordinated are triggered. Moreover, a filter on one visualization may result on a filter on another visualization. This kind of events are also handled.

To turn the actual abstracts' values into **visualizations**, the Gisplay (Cardoso et al. 2017) and the Highcharts Javascripts APIs were used. Other visualizations like the matrix plots were implemented. The interface is composed of three main areas as displayed In Figure 6.3:

1. **Global Abstracts Area (Figure 6.3-1)**: Global Abstracts, Compact Spatial Abstracts and Compact Temporal Abstracts are being coded into matrix plots. There is a matrix plot for each abstract. In our implementation, the value of these abstracts are real-values (as opposed to matrices or vectors). This way, one cell shows the value of an *Abstract* in a certain spatiotemporal LoD. One matrix shows the value of an *Abstract* for all the available LoDs. Blue shows low values while the green and yellow ones show high values. In the rows, we have the spatial granularities declared as valid for the argument *space* (bottom-up: from finer to coarser granularities), and in the columns, we have the temporal granularities declared as valid for the

argument *time* (left-right: from finer to coarser granularities). The skeleton of a matrix plot is displayed in Figure 6.2.

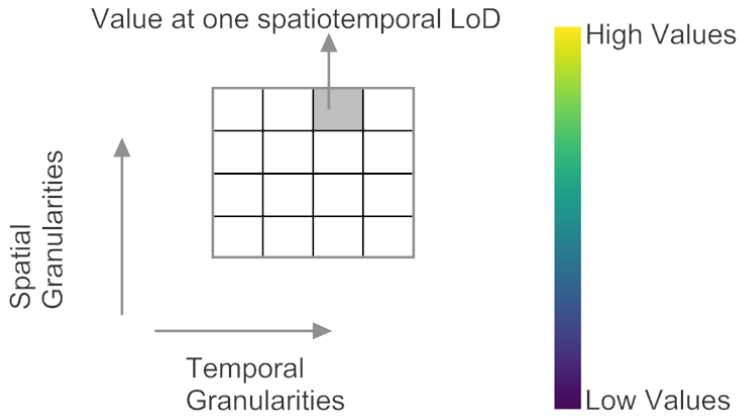


Figure 6.2: An overview of the structure of a matrix plot.

Recall that, the Compact Spatial Abstracts are abstracts computed from the Spatial Abstracts whereas Compact Temporal Abstracts are the abstracts computed from the Temporal Abstracts. Therefore, the symbol ▲ points out a Compact Spatial Abstract while the symbol ◷ indicates a Compact Temporal Abstract. When none of these icons is present it means that we are before a global abstract.

For example, in Figure 6.3, there are three matrices highlighted. The matrix *c* displays the Global Abstract - Occupation rate. The matrix *a* displays the Compact Spatial Abstract - Average of Spatial occupation rate. Finally, the matrix *b* shows the Compact Temporal Abstract - Average of Temporal occupation rate.

2. **Dynamic Abstract Area (Figure 6.3-2)**: This area can be used to show Global Abstracts, Spatial Abstracts and Temporal Abstracts. In Figure 6.3, the same Global Abstracts that are in the Global Abstracts Area are being displayed but using a Parallel Coordinate. Each line corresponds to one spatiotemporal LoD, and each coordinate corresponds to an Abstract.

   When a matrix has the symbol ▲ it indicates a Compact Spatial Abstract. As mentioned, the Dynamic Abstract Area can also be used to show other types of abstracts, namely Spatial Abstracts. In Figure 6.4, two types of Spatial Abstracts are being displayed. On the left side, the Spatial occupation rate and on the right-side, the Spatial collision rate. The matrix highlighted as *a* (see Figure 6.4) refers the Compact Spatial Abstract - Average of Spatial occupation rate. Therefore, one cell in that matrix (i.e., one spatiotemporal LoD) can be detailed in one Spatial Abstract (i.e, one time-series). Because there are four cells highlighted, there are four Spatial Abstracts displayed on the left-side of the Dynamic Abstract Area.

When a matrix has the symbol it indicates a Compact Temporal Abstract. For example, in Figure 6.5, the matrix referred as *a* shows the Compact Temporal Abstract - Average of Temporal occupation rate. Therefore, one cell in that matrix (i.e., one spatiotemporal LoD) can be detailed into one Temporal Abstract (i.e, one thematic map). Because there are two cells highlighted, there are two Temporal Abstracts displayed in the Dynamic Abstract Area. Notice that, when the spatiotemporal LoD has a *raster* granularity, the map represents each spatial granule through a point, leading to a dot map (e.g., the map on the right side). Otherwise, the spatial granules are displayed in their original form which leads to a Choropleth map (e.g., the map on the left side).

3. **Phenomena Representation (Figure 6.3-3)**: This area is used to display spatiotemporal events at a particular spatiotemporal LoD. The slider underneath allows to scroll temporally through the data, for each temporal granule, according to the spatiotemporal LoD that was chosen. Therefore, on the map, the number of events for each spatiotemporal granule is displayed through a thematic map. A dot map when the spatiotemporal LoD contains a *raster* granularity is displayed or a Choropleth map, otherwise. For example, in Figure 6.3, the Phenomena Representation area is displaying the data at the spatiotemporal LoD - *Counties*, *Weeks*.

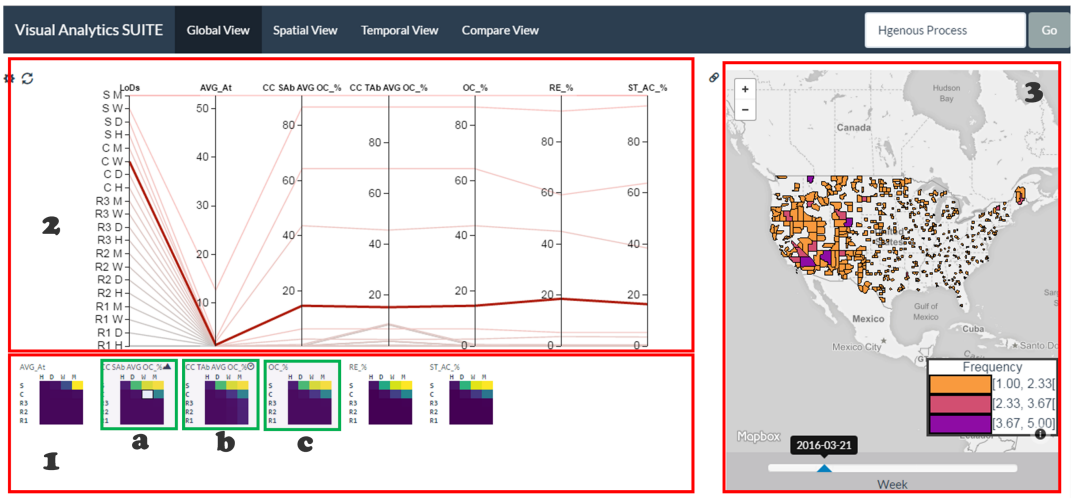


Figure 6.3: An overview of the SUITE prototype's interface.

SUITE is designed to help users in the detection and analysis of patterns within spatiotemporal events at multiple spatiotemporal LoDs. Our interface is thus composed of three main areas as presented. These areas follow a coordinated strategy among the visualizations provided. Coordinated views have been used to facilitate visualization (Weaver 2010). This method encourages understanding by facilitating data exploration through linked visualizations via user interaction. That is, the visualizations are not being used only to show information but also to serve as an interaction mechanism with other

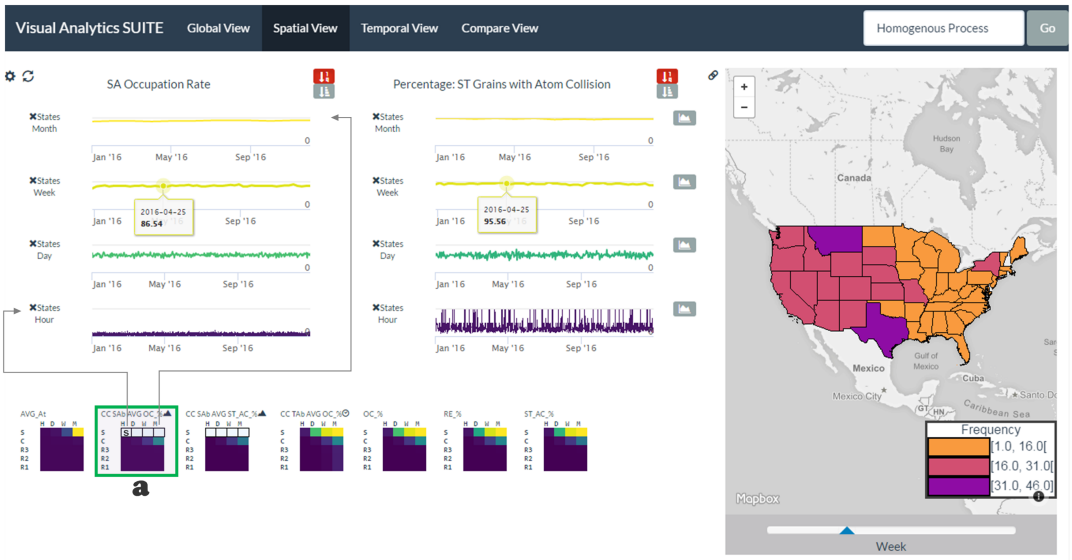


Figure 6.4: An overview of the SUITE prototype's interface with Spatial Abstracts.

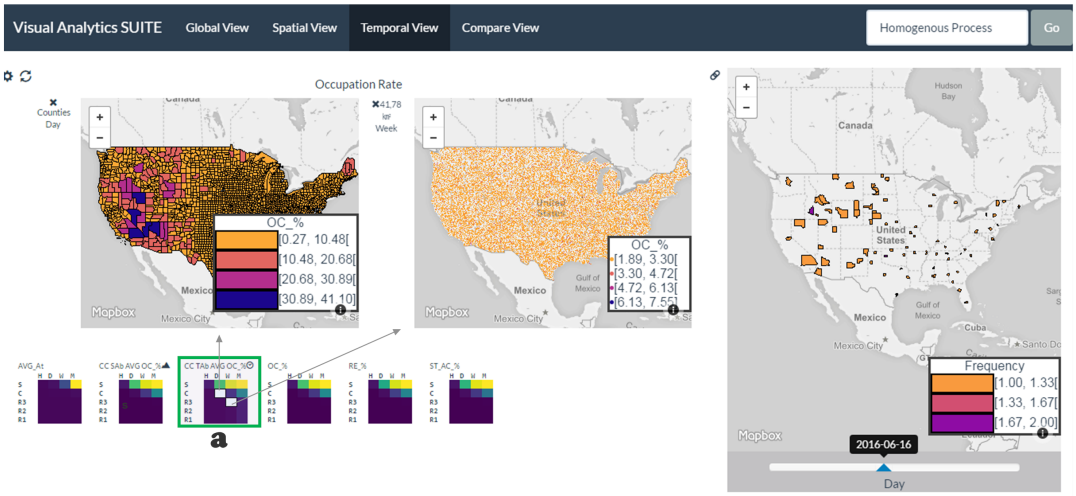


Figure 6.5: An overview of the SUITE prototype's interface with Temporal Abstracts.

views. Some videos regarding the SUITE-VA are available at `http://staresearch.net/ricardo-silva-may-2017/`.

The design of SUITE is intended to follow the VA Mantra: *"Analyze first, show the important, zoom, filter and analyze further, details on demand"* (Keim et al. 2008). First of all, the abstracts are precomputed and then, the interface starts by displaying global abstracts across spatiotemporal LoDs that may provide hints about different facets or patterns within spatiotemporal events. Then, one can analyze further by looking at Spatial or Temporal Abstracts. Finally, at any moment of the analyzes, one can check the reason behind some abstracts' values by visually inspecting the actual representation of the phenomenon at a particular LoD.

The prototype presented was used on synthetic and real datasets. The Abstracts

used to explore such datasets were presented in Section 5.5. The entire list of abstracts implemented can be seen in Appendix C.

6.2 Experiments on Synthetic Datasets

To produce synthetic datasets of spatiotemporal events, a configurable generator of spatiotemporal events was used. developed by (Gabriel et al. 2013) - R package (stpp). Using it, synthetic datasets were generated.

Stpp package allows us to simulate spatiotemporal point processes which in practice means spatiotemporal events where the event's spatial shape is a point. The spatiotemporal point processes are generated within a polygon and a single closed interval.

Gabriel et al. 2013 exposes a set of functions in order to simulate spatiotemporal events following different models (Møller and Ghorbani 2010; Gabriel et al. 2013; Gabriel 2014):

1. **Homogeneous Poisson Process**: the homogeneous Poisson process is the simplest mechanism for the simulation of a spatiotemporal point pattern. This model hardly approaches a pattern in a phenomenon but provides a good basis for comparison as it reflects complete spatiotemporal randomness. Informally, in a homogenous Poisson process, the events form an independent random sample from the uniform distribution on the spatiotemporal domain in which the events were simulated.

2. **Poisson Cluster Process**: the Poisson cluster process simulates spatiotemporal clusters of events. This model might reflect phenomena such as forest fires where several wildfire occurrences appear close in time and space, or the presence of spatiotemporal hotspots of crimes, for instance. Informally, a set of parents are generated, and afterward, a set of events are generated around each simulated parent. The dispersion of events in space and in time around each parent event is an input parameter through which we specify the spatiotemporal LoD. In this process, when events happen they occur near to each other in space and time. However, it is possible that no events occur.

3. **Contagious Process**: A contagious process can be pictured out as a cloud of events moving in space throughout time. the contagion process of a disease, for example, in which the disease is transmitted to other people through direct contact with an infected person. Informally, an initial event is generated, and afterward, the next events are generated near to locations of the previous event(s) simulated. The spatial and temporal neighboorhoods on which the next events are generated are input parameters through which we specify the spatiotemporal LoD.

4. **Log-Gaussian Cox Process**: The Log-Gaussian Cox process simulates spatiotemporal events such that some regions reveal higher intensity. This model might reflect

phenomena that contain geographic regions of higher risk, which might changes slowly over time. This pattern might happen with wildfires, infectious diseases, among others. Informally, the Log-Gaussian Cox process is a inhomogeneous Poisson process with a stochastic (i.e., randomly determined) intensity. In this case, we have no precise control of the spatiotemporal LoD in which the patterm is simulated.

Different datasets were produced following one or more of the models presented. The set of datasets simulated are displayed in Table 6.1 along with their characteristics like the model used to generate it, the number of events in it, and last but not least, the spatiotemporal LoD in which the pattern/model was simulated[1]. All the datasets were generated within the region of the USA and during one year.

The datasets were modelled using the granularities-based model, and then, the SUITE prototype was used to make analyses over them based on the Abstracts detailed in Section 5.5. The results are reported below.

6.2.1 Poisson Cluster Process

Let's start by the Dataset 2. This dataset was simulated with the Poisson Cluster process and is composed by 30.000 events within the region of the USA that occurred during one year. The clusters of events are built around a parent within a spatial distance of 110 km and a temporal distance of one day.

The dataset was modeled through a synthetic predicate, with two arguments $synthetic(space, time)$. The granular term required to model these events was only the identity function symbol.

Regarding Dataset 2, the most detailed spatial granularity $Raster\left(0.16km^2\right)$ is based on grid of 16384 x 16384 cells that cover the analyzed spatial extent of the phenomenon, and each cell has an area of 0.16 km^2. The coarser spatial granularities were obtained by dividing by a factor of 4 the number of cells in the grid. So the valid granularities for

[1] The code to simulate the datasets and the actual datasets are available in the repository `http://github.com/RFASilva/SimulatedDataSets`

Table 6.1: Datasets of spatiotemporal events simulated.

	Model	Number Events	LoD
Dataset 1	Homogenous	30.000	NA
Dataset 2	Poisson Cluster	30.000	110 Km, Day
Dataset 3	Poisson Cluster	30.000	2 Km, Week
Dataset 4	Poisson Cluster + Homogenous	33.000	110 Km, Day
Dataset 5	Poisson Cluster	30.000	570 Km, Week
Dataset 6	Contagious	5.000	110 Km, Week
Dataset 7	Log-Gaussian Cox	15.000	NA

space were rasters with cell sizes approximately of 0.16 km^2, 2.55 km^2, 41.74 km^2. The granularities *Counties* and *States* were also included. The time granularities used were *Hours*, *Days*, *Weeks*, *Months*.

The raw data (events) were encoded at the base LoD of the *synthetic* predicate, which includes the time granularity *Hours* and the space granularity $Raster\left(0.16km^2\right)$. After that, the granularities-based module was used to automatically produce the data for all LoDs of the *synthetic* predicate and the SUITE module was used to precompute all the abstracts defined for each LoD. Using the Interface module, our analyses started by looking at global abstracts. Figure 6.6 shows the global abstracts (i.e., the Occupation rate, the Collision rate and the GMBN) for all the spatiotemporal LoDs of Dataset 2. For the sake of simplification, spatiotemporal LoD will be written as LoD_{st} in the following sections.

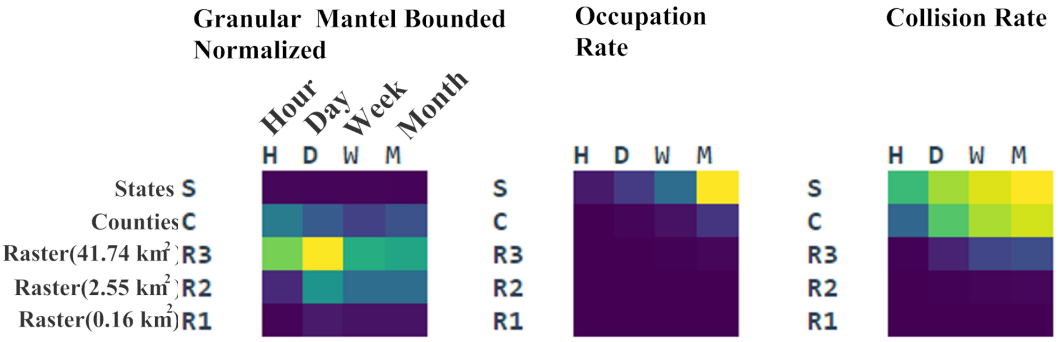


Figure 6.6: Global Abstracts: *GMBN*, Occupation rate and Collision rate describing Dataset 2.

The *GMBN* points to the LoD_{st} - ($Raster(41.74km^2)$, *Days*) as the one with greatest spatiotemporal interaction. This seems to be compliant with the LoD_{st} in which the pattern was simulated. Regarding the other global abstracts (i.e., the Occupation rate, and the Collision rate), their values increases as long as we move to coarser $LoDs_{st}$. This happens because as long as we move to coarser $LoDs_{st}$, the co-occurrence of granular syntheses in spatiotemporal granules increase, once the number of spatiotemporal granules available at coarser $LoDs_{st}$ decreases. Nevertheless, according to the phenomenon, the values of Occupation rate and Collision rate might increase at different rates.

In order to better understand in what $LoDs_{st}$ the perception of the phenomenon distinguishes itself, we use an instrument from the interface module that allows us to correlate two global abstracts.

We have implemented two forms of observing the correlation between two global abstracts. One of them is called correlation evolution through spatial granularities, which allows to observe for each spatial granularity how the correlation behaves, considering all the temporal granularities. The other is called correlation evolution through temporal granularities, which allows us to observe for each temporal granularity how the correlation behaves with respect to all the spatial granularities.

Figure 6.7a illustrates the correlation evolution through spatial granularities between the GMBN and the Collision rate. Each spatial granularity gives origin to a series in the chart. On the other hand, Figure 6.7b illustrates the correlation evolution through temporal granularities between the GMBN and the Collision rate. In this case, each temporal granularity gives origin to a series in the chart. The color encodes the spatial granularity while the shape of the markers encodes the temporal granularity. This encoding scheme is the same on both forms of correlation. Therefore, a marker with a particular color and shape represents the same spatiotemporal LoD on both charts.

Moreover, in the correlation evolution through spatial granularities the lines connect markers with the same color (i.e., the spatial granularity is the same) while in the correlation evolution through temporal granularities the lines connect markers with the same shape (i.e., the temporal granularity is the same). Notice that, both charts might become cluttered according to the data that are being mapped. To attenuate that problem, a user can hide or make visible series of the chart interacting with the legend.

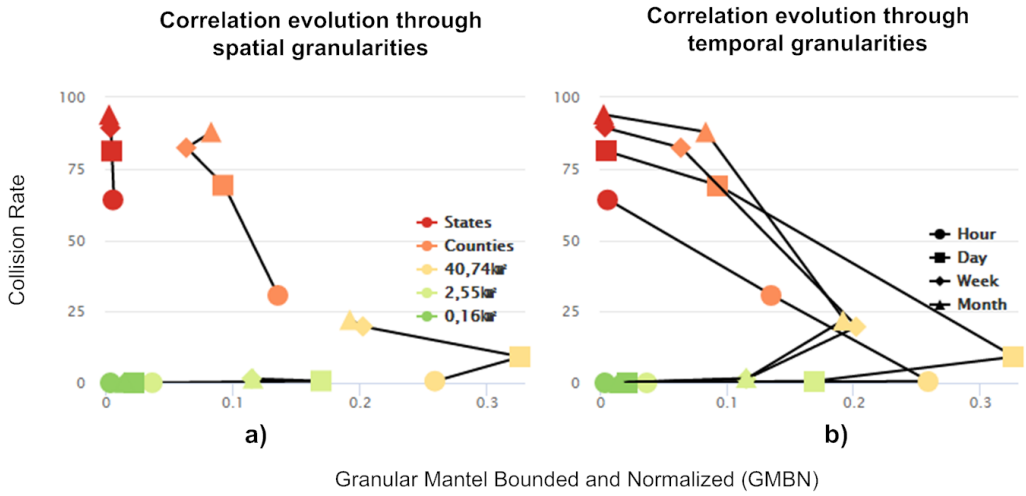


Figure 6.7: Correlation between the GMBN and the Collision rate.

On both charts we can observe "elbows". An elbow tip, in these charts, has a particularity that it might be interesting to explore. For the discussion that follows, let's assume that an elbow is created by going from a finer granularity to a coarser granularity (e.g., as happens in the series regarding the spatial granularity $Raster(40,74km^2)$ in Figure 6.7a. In these cases, it seems that there is a granularity G such that: (i) for granularities finer than G the correlation seems to be positive; (ii) for granularities coarser than G the correlation seems to be negative. This might be a hint about the LoDs in which the perception of a phenomenon distinguishes itself, considering the two global abstracts at study.

In Figure 6.7a, an elbow is visible taking into account the spatial granularity $Raster(40,74km^2)$, where the elbow tip is reached at the granularity *Days*. In Figure 6.7b, the elbow most pronounced is revealed at the temporal granularity *Days* where the

elbow tip is reached at the granularity $Raster(40,74km^2)$.

Therefore, the LoD_{st} - ($Raster(40,74km^2)$, Days) is where the elbow tip is observed on both charts. This conclusion is similar to the one achieved by just looking at the GMBN, in Figure 6.6, and this analysis might seem useless. However, looking at only one Global Abstract as a way of understanding suitable $LoDs_{st}$ to detail our analyses might be misleading. These scenarios will be discussed later.

The correlation between the GMBN and the Collision rate serves two purposes. First, there is one more hint pointing to ($Raster(40,74km^2)$, Day) as a suitable LoD_{st} to analyze the data. Second, it allows us to introduce the correlation charts.

Given the evidences pointing that there might be a pattern in the LoD_{st} - ($Raster(40,74km^2)$, Day), or at least the phenomenon is observable in such LoD_{st}, we use the *Phenomenon Representation* area to have a grasp of the data at such LoD_{st}. The data at three different temporal granules chosen without any particular criterion are displayed in Figure 6.8. As you can see, there are clusters of events happening over the USA.

Figure 6.8: Dataset 2 at the spatiotemporal LoD $Raster(41.77\ km^2)$ and *Days* displayed in three temporal granules.

The analysis made so far points out that the Dataset 2 might have a spatiotemporal pattern and such pattern might be better perceived at $(Raster(40,74km^2), Days)$. The pattern in question are clusters of events happening over time.

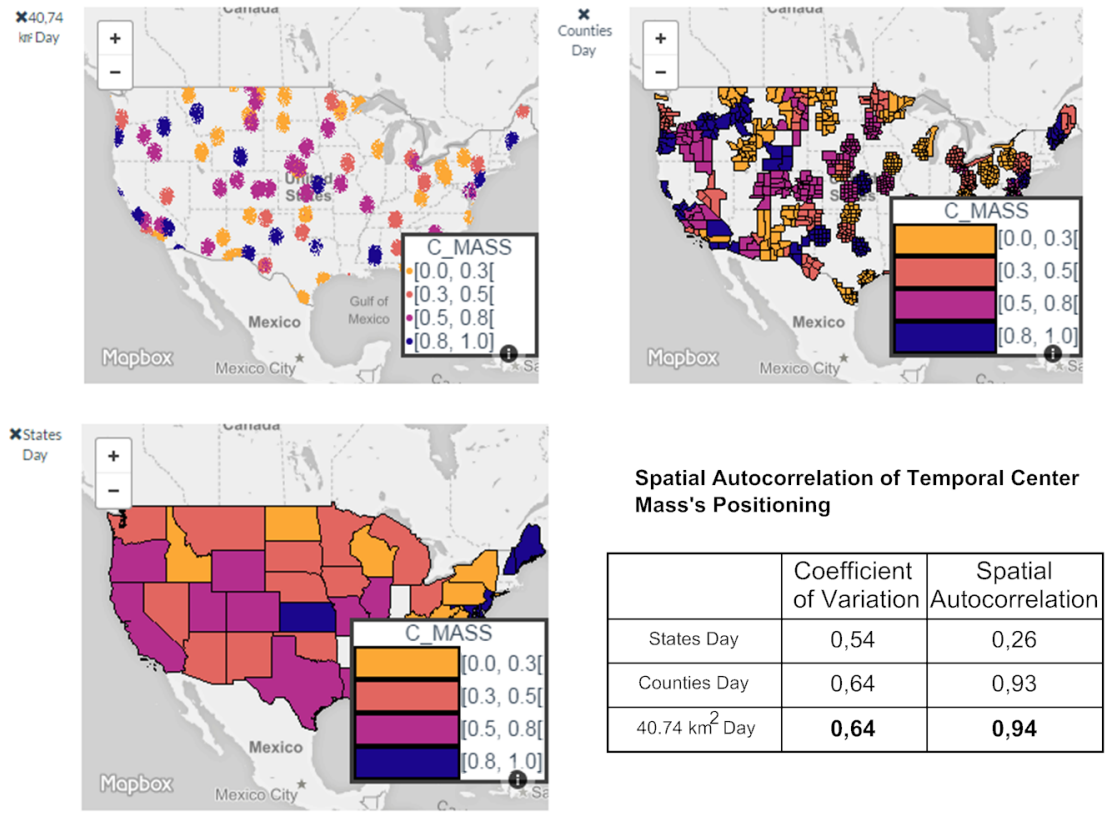
Our analyses were further detailed using the Spatial and the Temporal Abstracts in order to confirm a pattern in the LoD_{st} - $(Raster(40,74km^2), Days)$.

We start by looking to the Temporal Abstract - Temporal Center Mass's Positioning for three $LoDs_{st}$ as can be seen in Figure 6.9. The $LoDs_{st}$ are: $(Raster(40,74km^2), Days)$ $(Counties, Days)$ and $(States, Days)$. Orange means that most of the events that occurred in the spatial granule were old while dark blue means that most of the events occurred in the spatial granule were recent in what concerns the extent of the temporal granularity.

Looking at the $LoDs_{st}$ - $(Raster(40,74km^2), Days)$ and $(Counties, Days)$, in Figure 6.9,

the geographic regions where the clusters of events have happened can be identified, since spatial granules close to each other have similar values of the Temporal center mass's positioning. In other words, the events occurring near in space seems to occur near in time.

The previous conclusions are also captured by the two Compact Temporal Abstracts of the Temporal center mass's positioning, i.e., its Coefficient of Variation and its Spatial autocorrelation. In this case, the coefficient of variation tells us in what $LoDs_{st}$ the value of the Temporal Center Mass's Positioning varies more among the spatial granules while the Spatial autocorrelation measures how the value of the Temporal Center Mass's Positioning is similar in neighboring spatial granules. Thus, we are interested in $LoDs_{st}$ such that there is a considerable variation and the spatial autocorrelation's value suggests spatial correlation. In what concerns the three $LoDs_{st}$ displayed in Figure 6.9, the $LoDs_{st}$ - $(Raster(40,74km^2), Days)$ is where the Coefficient of Variation and the Spatial autocorrelation take the highest values as detailed in Figure 6.9. The spatial autocorrelation is 0.94 (strong positive correlation) and the coefficient of variation is 0.64. Clusters are spread out across the entire USA. Besides that, we can relate the geographic regions and the time moments in which the clusters occurred. This kind of perception is lost if you look at the data in the $LoDs_{st}$ - $(States, Days)$ (see Figure 6.9), for example.



Spatial Autocorrelation of Temporal Center Mass's Positioning

	Coefficient of Variation	Spatial Autocorrelation
States Day	0,54	0,26
Counties Day	0,64	0,93
40.74 km² Day	**0,64**	**0,94**

Figure 6.9: The Temporal Center Mass's Positioning for three $LoDs_{st}$.

Since clusters are happening over time, we use the Compact Spatial Abstract - Spatial

average nearest neighbor (Spatial ANN) and its z-score in order to understand when those clusters of events are happening.

Four $LoDs_{st}$ were chosen: $(Raster(40,74km^2), Hours)$, $(Raster(40,74km^2), Days)$, $(Raster(40,74km^2), Week)$, $(Raster(40,74km^2), Month)$. These were chosen because we know, based on evidence, that the $LoDs_{st}$ - $(Raster(40,74km^2), Days)$ is appropriate to analyze the data. So, the $LoDs_{st}$ - $(Raster(40,74km^2), Days)$ is included in the next analysis. This leaves us with the possibility of varying the spatial or the temporal granularity. But the previous analysis allows to note that the spatial granularity $Raster(40,74km^2)$ was able to show the places where the clusters happened. For this reason, we vary the temporal granularity.

The Spatial Abstracts are displayed in Figure 6.10. Notice that, the set of time series for each Temporal Abstract share the extremes of the Y axes. Besides that, the color of a time series is given by the color used on the corresponding Compact Spatial Abstract (i.e., matrix plot).

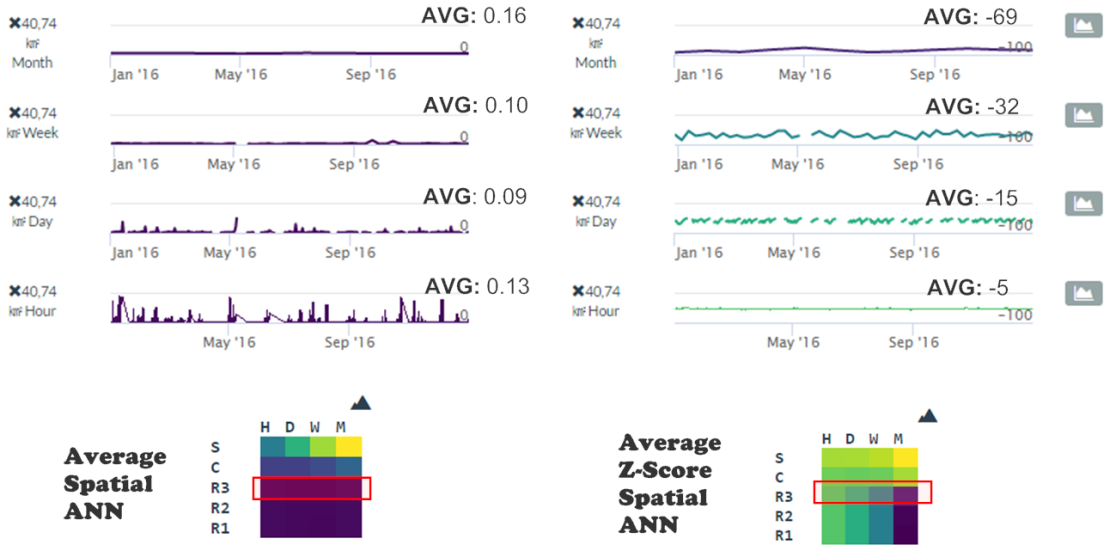


Figure 6.10: The Spatial Average nearest neighbor and its z-score in four $LoDs_{st}$.

Recall that, if the value of the Spatial ANN is less than 1, the trend is toward spatial clustering while if the value is greater than 1, the trend is toward dispersion. Very low or very high z-score values suggest some spatial pattern, and therefore, we can reject the complete spatial randomness.

Based on Figure 6.10, the Spatial Abstracts revealed a clustered phenomenon over time, since the average of the Spatial ANN values points to clusters of events throughout time. In the LoD_{st} - $(Raster(40,74km^2), Hours)$ we can observe variations between a clustered and a non clustered phenomenon. But in the remaining $LoDs_{st}$, the phenomenon reveals to be quite stable and clustered because the values of the Spatial ANN are constantly close to zero and the corresponding z-scores are quite negative (i.e., the z-score is not close to zero).

As these two Spatial Abstracts complement each other, we plot them in a scatter plot, using the interface (a click on the right-side buttons displayed in Figure 6.10). These scatter plots are displayed in Figure 6.11. Notice that, the extremes on both axes are relative to the LoD_{st} shown.

Each point in a scatter plot shows the values of the two Spatial Abstracts that occurred at a particular temporal granule. Therefore, the number of points in a scatter plot is equal to the number of temporal granules in the temporal granularity that composes the LoD_{st} being displayed.

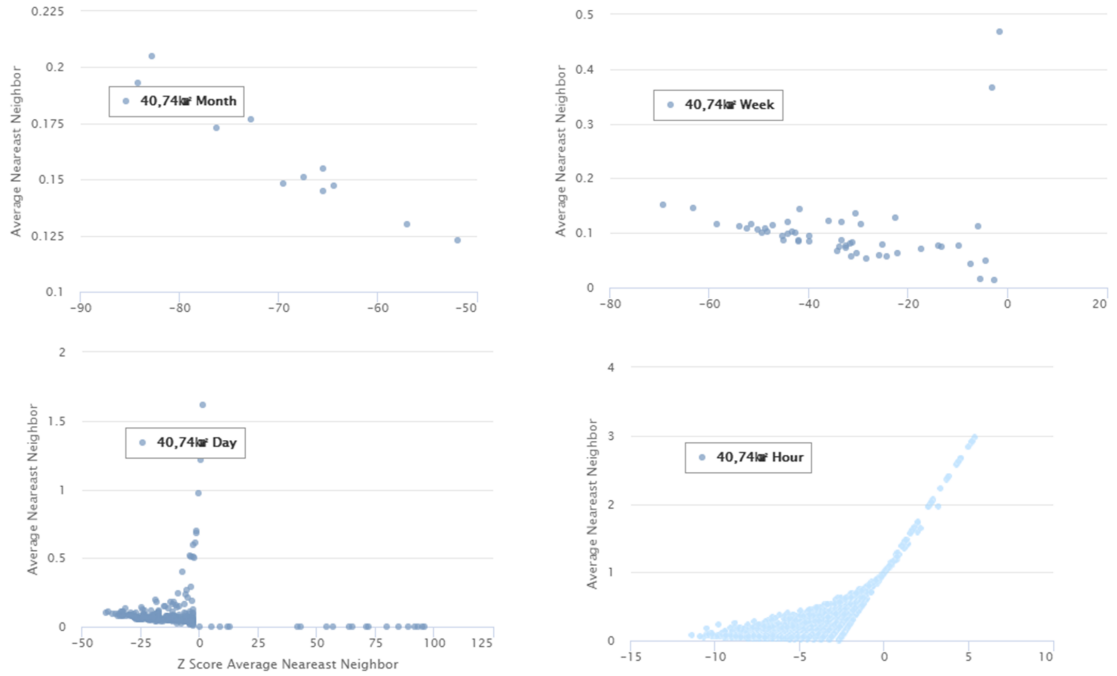


Figure 6.11: The Spatial ANN and its Z-score displayed in four $LoDs_{st}$.

At the LoD_{st} - $(Raster(40,74km^2), Hours)$ - bottom-right, there are many points holding a value close to zero of the Spatial ANN, and their z-scores are not so negative as the ones in the others $LoDs_{st}$. Looking at the LoD_{st} - $(Raster(40,74km^2), Month)$, it seems that the phenomenon is always clustered (i.e., the values of Spatial ANN are close to zero and their z-scores quite negative). Finally, regarding the LoD_{st} - $(Raster(40,74km^2), Day)$ and LoD_{st} - $(Raster(40,74km^2), Weeks)$ seems to be the $LoDs_{st}$ that better fit the Poisson Cluster process. Recall that, in a Poisson Cluster process, events occur near other events but there are a few times where no events occur. This is visible in the scatter plots of the LoD_{st} - $(Raster(40,74km^2), Day)$ and LoD_{st} - $(Raster(40,74km^2), Weeks)$, once the majority of the points have the values of the Spatial ANN close to zero and their z-scores are quite negative. However, there are also points where the values of the Spatial ANN are close to zero and their z-scores are positive (no clustering) and also there are points with values of the Spatial ANN that are far from zero (no clustering).

In short, the analysis made over Dataset 2 that contains a Poisson cluster process

simulated with clusters of events dispersed within 110 km and one day around their parents was:

- We use the matrix plots to analyze the GMBN, Occupation rate and Collision rate. Here, the GMBN pointed to the LoD_{st} - $(Raster(40,74km^2), Days)$
- We correlate the GMBN and Collision rate using the correlation of evolution through spatial granularities and through temporal granularities. Again, the LoD_{st} - $(Raster(40,74km^2), Days)$ was suggested.
- We used the phenomenon representation area to have an overview of the phenomenon at LoD_{st} - $(Raster(40,74km^2), Days)$ in three temporal granules chosen without any particular criterion. Clusters of events were observed.
- The Temporal Abstract - Temporal Center Mass's Positioning was studied in three different LoDs. Furthermore, two Compact Temporal Abstracts were also analyzed: Coefficient of variation and the spatial autocorrelation. Here, the LoD_{st} suggested was also LoD_{st} - $(Raster(40,74km^2), Days)$ if one wants to understand in what periods of time clusters of events occur in certain geographic regions. It was also possible to observe that the clusters are spread out over the entire area of the USA.
- The Spatial Abstracts - Spatial average nearest neighbor (Spatial ANN) and its z-score was used in order to understand not only when the clusters are happening but also what LoD_{st} better fits the Poisson Cluster process. The analysis suggested that clusters are distributed throughout the one "year" in which data was simulated. Finally, the analysis suggests that the LoD_{st} that better fits the Poisson Cluster process is LoD_{st} - $(Raster(40,74km^2), Days)$ or LoD_{st} - $(Raster(40,74km^2), Weeks)$.

Other datasets were simulated following the Poisson cluster model - the Datasets 3, 4 and 5. These datasets were simulated within the USA boundaries over a year. In **Dataset 3**, each cluster of events was built around a parent within a spatial distance of 2 km and a temporal distance of one week. **Dataset 4** is similar to Dataset 2 but contains an additional 3.000 events following a homogenous model. These 3.000 events are spread out over the same period of the 30.000 events that follow the Poisson Cluster model. Finally, in **Dataset 5**, each cluster of events was built around a parent within a spatial distance of 570 km and a temporal distance of one week. In the following analysis, we also add Dataset 1 that was simulated with the Homogeneous model.

The datasets described were also modeled using the *synthetic* predicate with similar valid granularities. All the granularities are equal with respect to the previous demonstration case except for the Raster granularities. This occurs because the mininum bounding box made by the events of the phenomenon might change from one dataset to another. Nevertheless, the most detailed spatial granularity is based on a grid of 16384 x 16384

cells and the other coarser spatial granularities were obtained by dividing the grid by a factor of 4.

Datasets 1, 3, 4, 5 will be discussed more briefly. That is to say, we will only discuss whether the SUITE-VA points to suitable LoDs to detail our analyses once the "detailed" analyses would be similar to the ones made over Dataset 2. Furthermore, a comparison between the abstracts' values obtained by a Poisson Cluster dataset or a Homogenous dataset is made.

Figure 6.12 shows the global abstracts for all spatiotemporal LoDs of datasets 1, 3, 4 and 5. First of all, the Occupation rate follows a similar pattern in all datasets. Dataset 3 stands out from the others regarding the Collision rate. This occurs because the clusters in Dataset 3 were simulated within a spatial distance of 2 km, and were thus much more spatially clustered than in the other datasets. As a result, the collision among granular syntheses starts to occur "sooner", i.e., in finer $LoDs_{st}$ when compared to the other datasets.

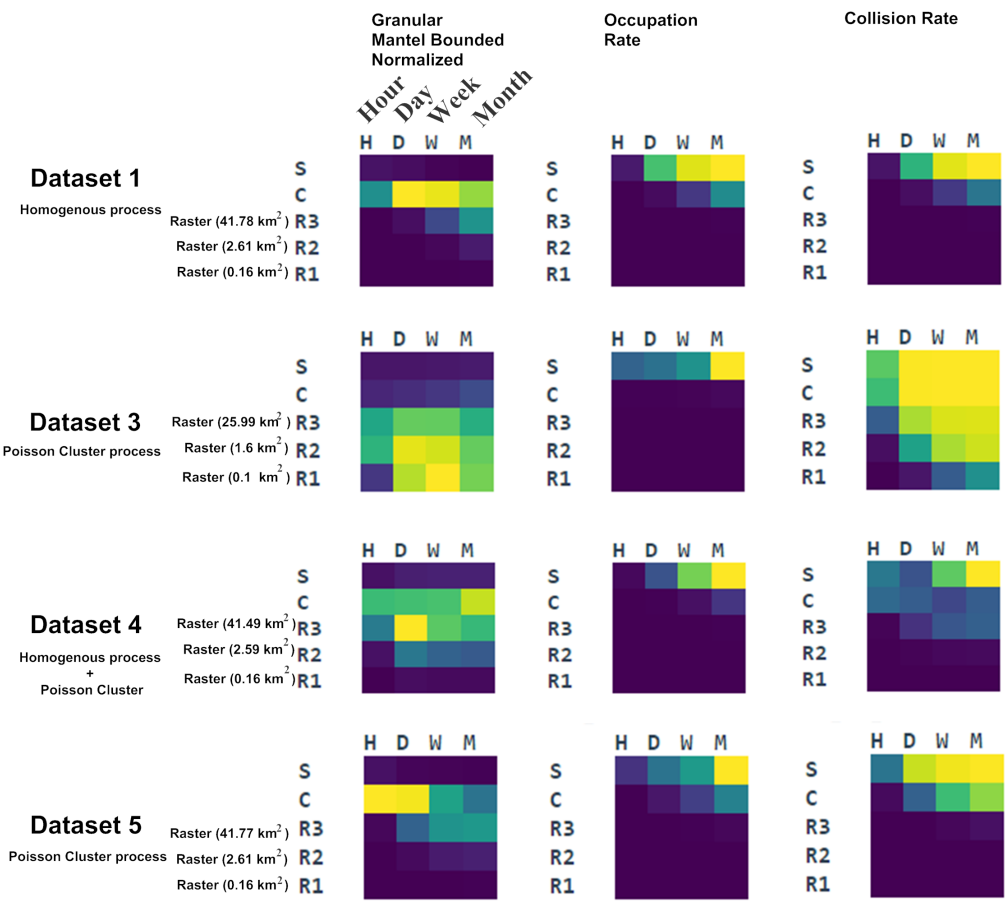


Figure 6.12: Global Abstracts regarding Dataset 1, 3, 4, 5.

As to Dataset 3, the GMBN highlights the following $LoDs_{st}$: (i) $(Raster(0.1km^2), Days)$; (ii) $(Raster(0.1km^2), Weeks)$; (iii) $(Raster(1.6km^2), Days)$; (iv) $(Raster(1.6km^2), Weeks)$. In this case, the values of spatiotemporal interaction are similar among the four $LoDs_{st}$,

and therefore, any of the $LoDs_{st}$ highlighted is potentially suitable to detail our analyzes. Nevertheless, the $LoDs_{st}$ - $(Raster(1.6km^2), Weeks)$ is the LoD_{st} that better approaches the LoD_{st} in which the data was simulated, once each cluster of events was simulated around a parent within a spatial distance of 2 km and a temporal distance of one week.

Dataset 4 is similar to Dataset 1, complemented by a homogenous process. In this case, the GMBN suggest the LoD_{st} - $(Raster(41.49km^2), Days)$, which is the LoD_{st} that better approaches the LoD_{st} in which the pattern is simulated, once each cluster of events was simulated around a parent within a spatial distance of 110 km and a temporal distance of one day.

Nevertheless, a single Global Abstract should not be used in order to immediately guide our analyses for one or more $LoDs_{st}$. So far, we have been using four global abstracts in order to have a grasp of the data. From these four abstracts, one is neighborhood dependent (GMBN) and the remaining ones are not (Occupation rate, Reduction rate) (check Section 5.3 for abstracts' properties). In other words, only the GMBN captures in their computation the spatiotemporal dynamics of events. Therefore, restricting ourselves to just one global abstract that looks for spatiotemporal patterns or properties of the spatiotemporal interaction might wrongly suggest one or more $LoDs_{st}$ as demonstrated below.

In Dataset 5, the GMBN highlights the following $LoDs_{st}$: (*i*) $(Counties, Hours)$; (*ii*) $(Counties, Days)$. However, each cluster of events was simulated around a parent within a spatial distance of 570 km and a temporal distance of one week.

The problem is that the events within a cluster are spatially "dispersed" (570 km) and the GMBN is not capable of capturing such situation. But even worse, in Dataset 1, the $LoDs_{st}$ - $(Counties, Days)$ and $(Counties, Weeks)$ are pointed as potential $LoDs_{st}$ in which there might be spatiotemporal interaction. However, this dataset was generated following a Homogenous model. This kind of scenarios can be easily discarded when we analyze several Global Abstracts that are looking for spatiotemporal patterns, or Global Abstracts with Compact Spatial Abstracts, or Global Abstracts with Compact Temporal Abstracts, or even all together.

To illustrate the previous idea, we analyzed the correlation between the GMBN and the Average of the Spatial ANN (Compact Spatial Abstract) for the different datasets as displayed in Figure 6.13.

Let's consider Dataset 1 that is the one with the Homogenous process. The correlation charts shows that when the GMBN reaches its maximum value, the value of the Average of the Spatial ANN is much greater than 1 (squared orange marker). Therefore, this phenomenon follows hardly a clustered pattern over time because in that case the value of the Average Spatial ANN would be closer to 0, something that did not happen in any LoD_{st} as the minimum value achieved was 0.8.

In Dataset 3, we have a clear hint about the $LoDs_{st}$ where the pattern was simulated because when the GMBN reaches its maximum value the Average of the Spatial ANN is close to zero (diamond green marker), as opposed to what happens in Dataset 1 (see

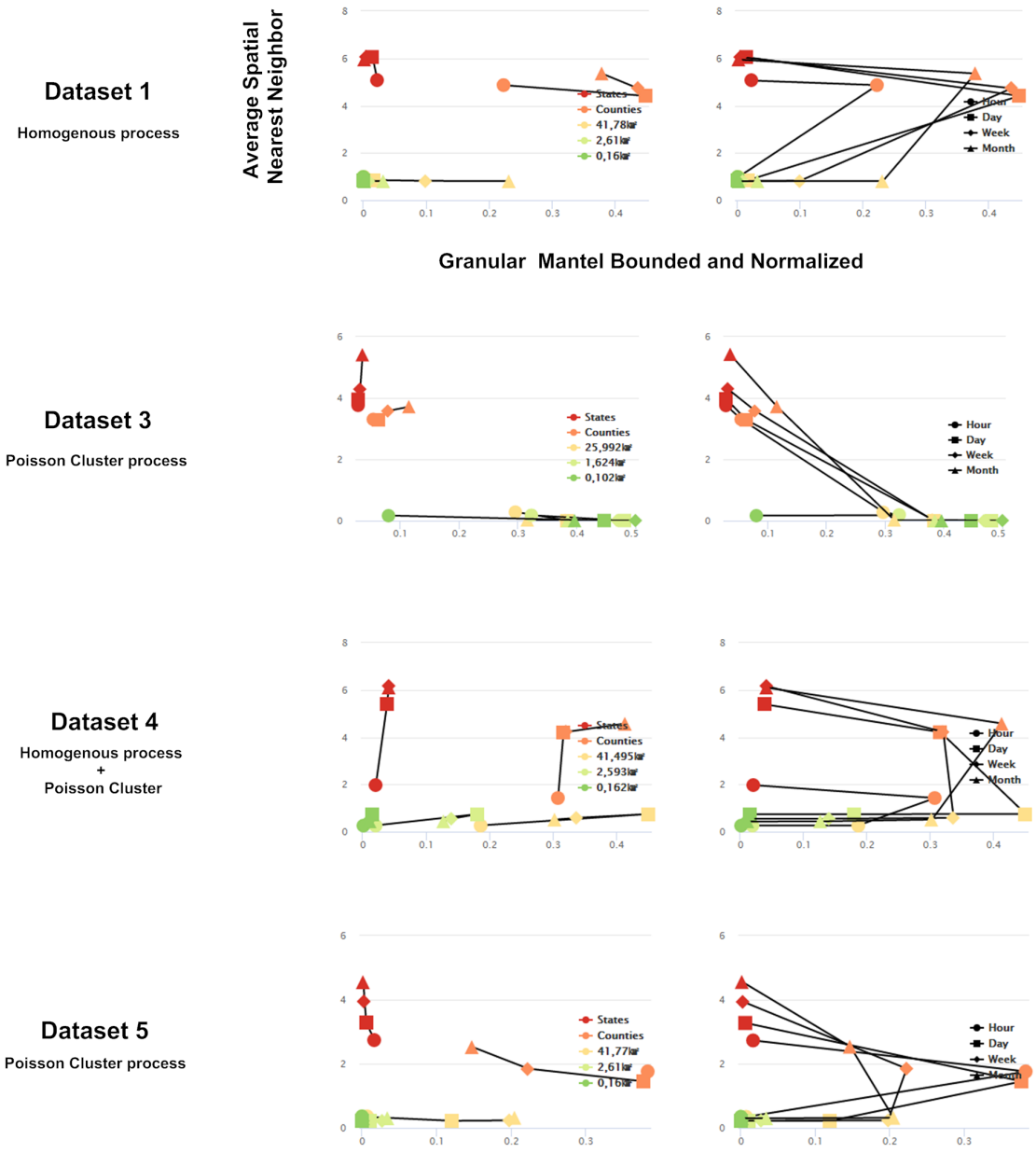


Figure 6.13: Correlation between the GMBN and the Average of the Spatial ANN (Compact Spatial Abstract).

Figure 6.13).

Looking at Dataset 4, the most pronounced elbow tip in the chart on the left (yellow square marker) corresponds to the LoD_{st} - $(Raster(41.49km^2), Days)$. This LoD_{st} is the one that better approaches the LoD_{st} in which the pattern was simulated, because each cluster of events was simulated around a parent within a spatial distance of 110 km and a temporal distance of one day. Despite the fact that GMBN reaches it's maximum in the yellow square LoD_{st} - $(Raster(41.49km^2), Days)$, the value for the Average of Spatial ANN is 0.5 which makes the hint weaker than in the case of Dataset 3. However, this gives us a clue for the right LoD_{st}.

Regarding Dataset 5, we do not have a clear hint about the $LoDs_{st}$ in which the data should be analyzed. Recall that, in this dataset, each cluster of events was simulated around a parent within a spatial distance of 570 km and a temporal distance of one week. So, the events are not that clustered. Therefore, the pattern is not so pronounced when compared with the other datasets. That being said, when the GMBN reaches its maximum value the Average of the Spatial ANN is **not** close to zero (square and circle orange markers - the LoD_{st} - $(Counties, Hours)$ and $(Counties, Days)$). This result has similarities with Dataset 1 - Homogeneous process. However, in this case, two elbow tips are observed (i.e., $LoDs_{st}$) that are not so pronounced but the Average Spatial ANN is close to zero. These correspond to the LoD_{st} - $(Raster(41.77km^2), Weeks)$ (i.e., the diamond yellow marker) and $(Raster(41.77km^2), Months)$ (i.e., the triangle yellow marker). In this case, the SUITE-VA provide a hint about two $LoDs_{st}$ such that one of them (i.e., LoD_{st} - $(Raster(41.77km^2), Weeks)$) may be appropriate to detail further analyzes.

The previous analysis would not be as clear for an user that is unfamiliar with the abstracts implemented as well as the interpretation of the visualizations provided. This relates to the learning curve concept. As a user is gaining more experience with the SUITE-VA, the understanding about the concepts involved will also become clearer.

A final remark about the interpretation of the correlation charts. The elbow tips provide a change from a positive to a negative (or vice-versa) correlation that might be interesting to explore. Nevertheless, there might be LoD_{st} of interest that do not correspond to elbow tips. Yet, according to the values that they hold for the abstracts at study, they might be also interesting to explore as in Dataset 5.

6.2.2 Contagious Process

Dataset 6 was simulated following the contagious process. The dataset was simulated within the USA boundaries over a year and is composed of 5.000 events. Based on an initial event, the next ones are generated within a spatial distance of 110 km and a temporal distance of a week. Furthermore, the dataset was modeled through the *synthetic* predicate. In this case, the most detailed spatial granularity $Raster\left(0.05km^2\right)$ is based on grid of 16384 x 16384 cells that cover the analyzed spatial extent of the phenomenon, and each cell has an area of 0.05 km^2. The other coarser spatial granularities were obtained

by dividing the number of cells in the grid by a factor of 4. So the valid granularities for space were rasters with cell sizes approximately of 0.05 km^2, 0.8 km^2, 12.5 km^2. The granularities *Counties* and *States* were also included. The time granularities used were *Hours*, *Days*, *Weeks*, *Months*.

To start our analysis we chose: (*i*) the GMBN; (*ii*) the Average of Spatial ANN; (*iii*) the Average of the z-score of the Spatial ANN; (*iv*) the Average of Temporal ANN; (*v*) the Average of the z-score of the Temporal ANN. The first three abstracts were already used so we skipped more explanations. In what concerns the Temporal ANN, for each spatial granule, in a given $LoDs_{st}$, how the granular syntheses are dispersed or clustered in time is measured as detailed in Section 5.5.

The Parallel Coordinates was used to simultaneously analyze the global abstracts chosen across all the $LoDs_{st}$. In this case, we are interested in understanding $LoDs_{st}$ in which (*i*) the phenomenon seems to be more clustered over time; (*ii*) the phenomenon seems to be more clustered over space; (*iii*) the $LoDs_{st}$ where the spatiotemporal interaction of events seems to be better perceived. To conduct such analysis, we filtered the Parallel Coordinates in each coordinate.

This way, interactively, we just considered $LoDs_{st}$ with values below 0.4 (approximately) regarding the average of the Spatial ANN. For the average of its z-score, we just considered values below -10 (approximately). Furthermore, values below 0.1 (approximately) with respect to the average of the temporal ANN were considered. For its z-score, we considered values below -1. Finally, the top three values of the GMBN were considered, which means values above 0.08. The results are displayed in Figure 6.14.

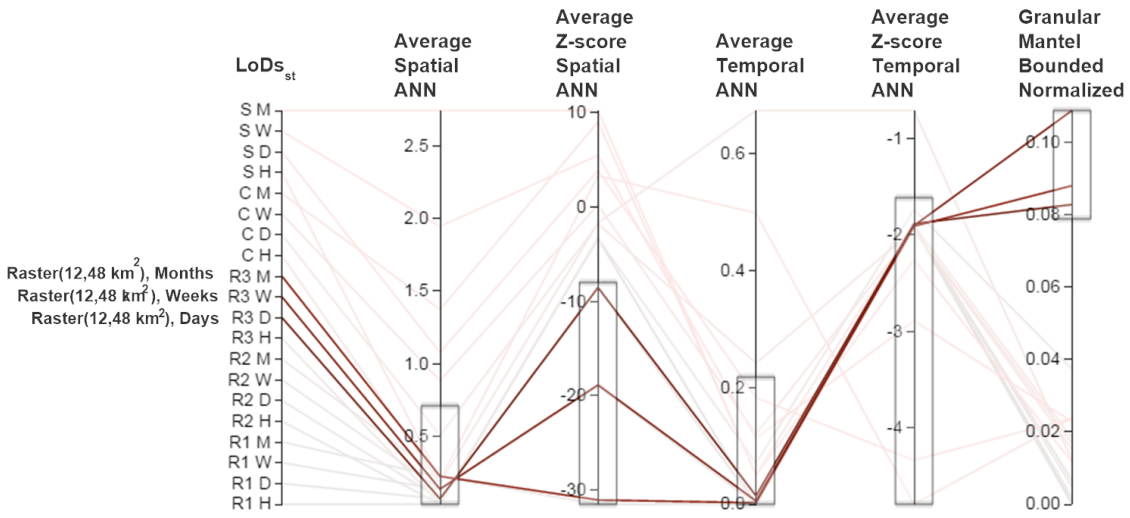


Figure 6.14: Overview about Dataset 6 using Global Abstracts, Compact Spatial Abstracts and Compact Temporal Abstracts.

Three $LoDs_{st}$ were highlighted: $(Raster(12.5km^2), Days)$, $(Raster(12.5km^2), Weeks)$ and $(Raster(12.5km^2), Months)$. Like it was done in Dataset 2, the Temporal Center Mass's Positioning was used in order to relate geographic regions with the center's of mass of

time at which events happened. This Temporal Abstract is displayed in Figure 6.16 for the three $LoDs_{st}$ identified. Regardless of the LoD_{st}, a grasp about the spatial path made

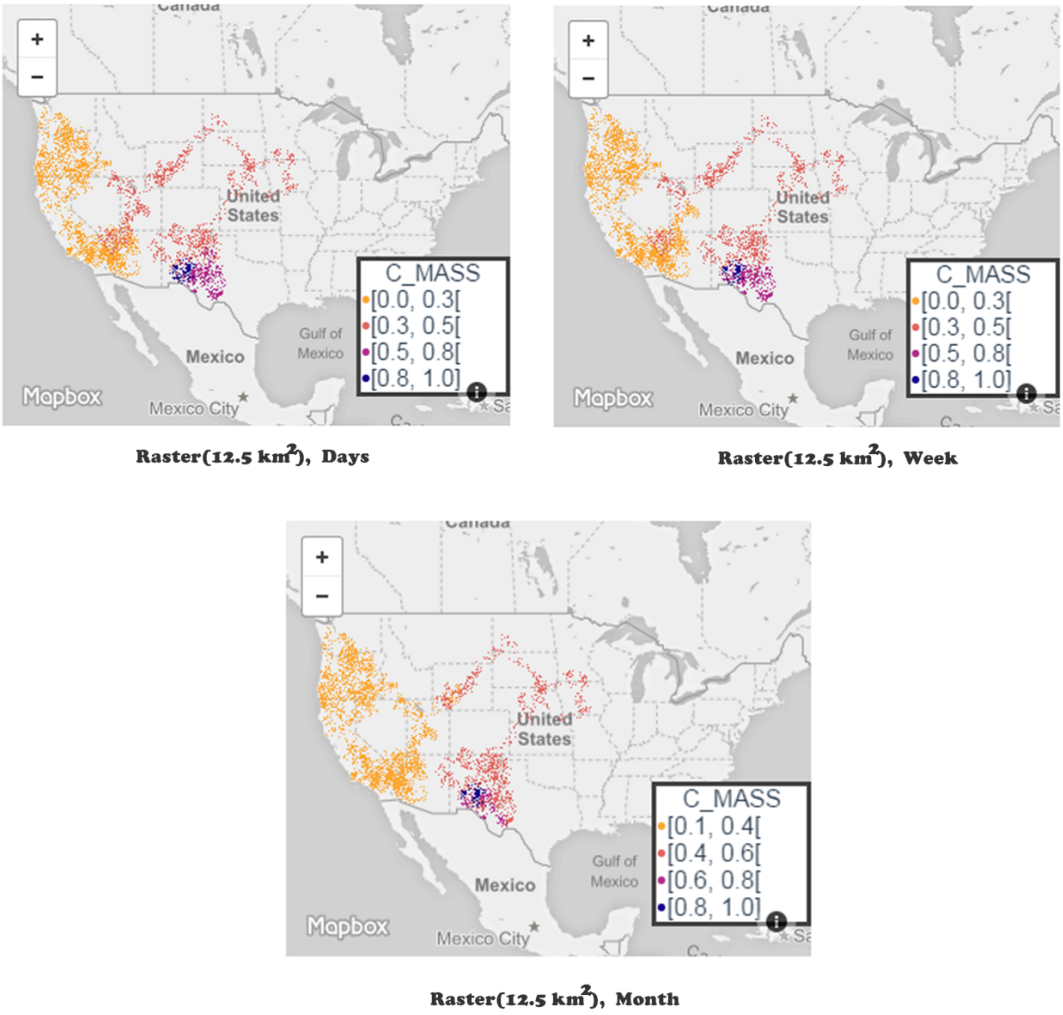


Figure 6.15: One Temporal Abstract at three different $LoDs_{st}$.

by the simulated contagious process is visible, thus confirming a contagious process. Nevertheless, in LoD_{st} - $(Raster(12.5km^2), Days)$ is where the path made is slightly better perceived.

Another experiment was made with two Spatial Abstracts: (*i*) the Spatial Scope; (*ii*) the Spatial Consecutive Distance between Centers of Mass. The former abstract indicates how much a phenomenon changes the size of its spatial extent over time while the latter measure whether such spatial extent moves in space over time. For the $LoDs_{st}$ identified initially, the Spatial Abstracts can be seen in Figure 6.16. Moreover, in the former abstract the average value is displayed while in the latter the coefficient of variation is shown.

Let's start by the Spatial Scope. In general, for the $LoDs_{st}$ identified, the phenomenon's spatial scope is quite stable throughout time with some variations here and there. However, the most stable LoD_{st} is the LoD_{st} - $(Raster(12.5km^2), Months)$.

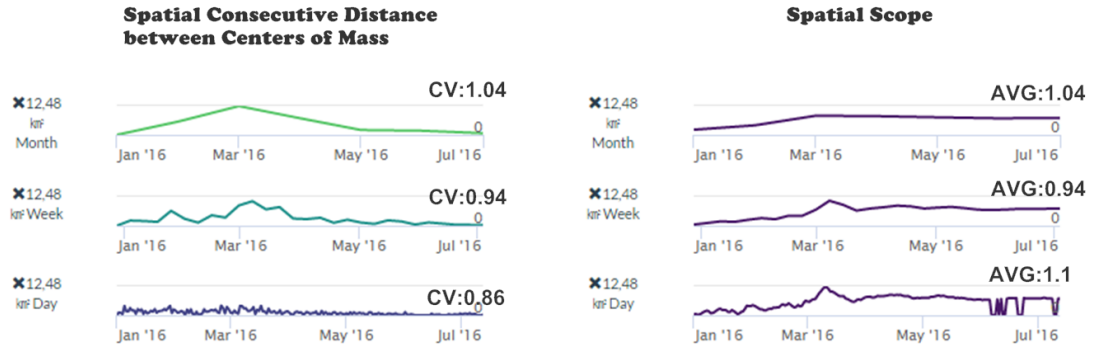


Figure 6.16: Two Spatial Abstracts about Dataset 6.

Regarding the Spatial Consecutive Distance between Centers of Mass, LoD_{st} - $(Raster(12.5km^2), Days)$ is where the distances between centers of mass seem to vary less according to the coefficient of variation. Thus, if one is interested in understanding how the contagious process evolved, in this simulated scenario, one should look at the LoD_{st} - $(Raster(12.5km^2), Days)$ because this is the LoD_{st} that seems to capture the smoothest transitions over time.

To conclude, in the Contagious process an initial event is generated, and then, the next events are simulated within a specified spatial and temporal distance. The dataset under analysis was generated with distances of 110 km and one week. The events generated within neighborhood are uniformly distributed and they are not necessary at a distance of a week. In fact, many of them might be at temporal distance less than a week. This might explain why, in the LoD_{st} - $(Raster(12.5km^2), Days)$, the Contagious process seems to be better perceived.

6.2.3 Log-Gaussian Cox Process

Dataset 7 was simulated following the Log-Gaussian cox process. The dataset was simulated within the USA boundaries over a year and is composed by 15.000 events. Therefore, this dataset will show geographic regions of higher incidence of events over others.

Furthermore, the dataset was modeled through the *synthetic* predicate. In this case, the most detailed spatial granularity $Raster\left(0.16km^2\right)$ is based on grid of 16384 x 16384 cells that cover the analyzed spatial extent of the phenomenon, and each cell has an area of 0.16 km^2. The remaining valid raster granularities for space were rasters with cell sizes approximately of 2.57 km^2, 41.27 km^2. The granularities $Counties$ and $States$ were also included. The time granularities used were $Days$, $Weeks$, $Months$.

As in Dataset 6 (Contagious process), we start by getting an overview of the set of the following abstracts using the Parallel Coordinates: (*i*) the GMBN; (*ii*) the average of Spatial ANN; (*iii*) the average of the z-score of the Spatial ANN; (*iv*) the average of temporal ANN; (*v*) the average of the z-score of the temporal ANN. Looking at the Parallel Coordinates:

There are no $LoDs_{st}$ holding values close to zero with respect to **Average Spatial ANN** containing quite negative z-scores. This kind of values suggest that we are not dealing with the Poisson cluster process as events occur close to each other in space.

There are some $LoDs_{st}$ holding values close to zero with respect to Average Temporal ANN but their z-scores are also close to zero. Also, for such cases, the spatiotemporal interaction is weak when compared with other $LoDs_{st}$. This kind of values suggests that we are not dealing with the Contagious process as events occur close to each other in space and in time.

Two $LoDs_{st}$ have the spatiotemporal interaction among events measured by the GMBN above 0.4, which is similar to the values obtained in Poisson Cluster simulated datasets. However, at this point, no particular meaning can be assigned to such values.

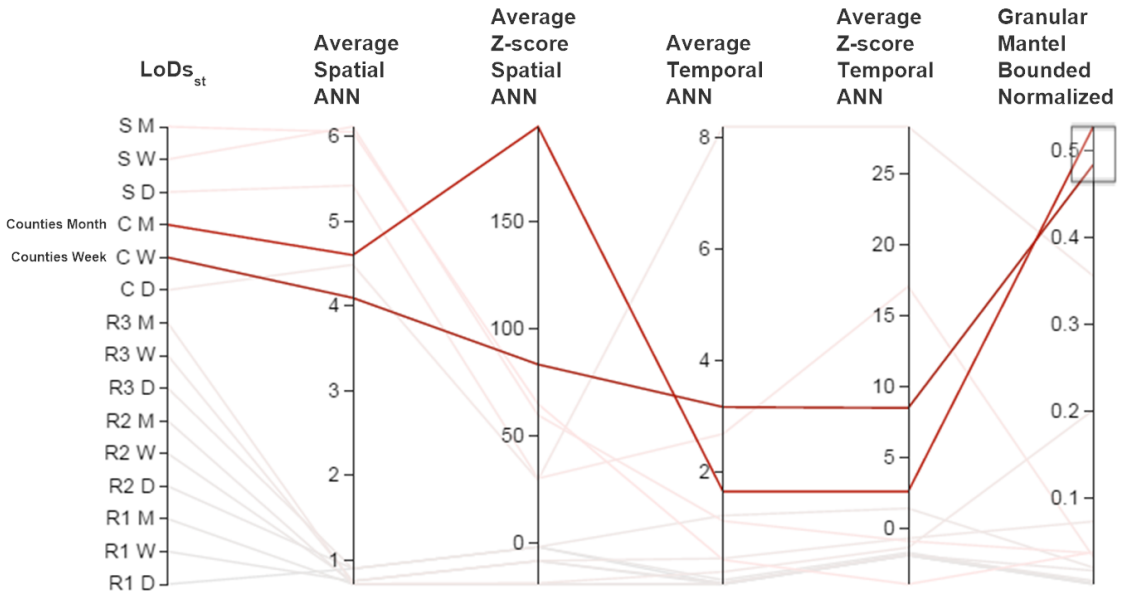


Figure 6.17: Overview about Dataset 7 (Log-Gaussian cox process) using Global Abstracts, Compact Spatial Abstracts and Compact Temporal Abstracts.

In Log-Gaussian Cox processes, we have geographic regions of higher incidence that might change slowly over time. This way, there are geographic regions that distinguish themselves from others in terms of the number of events that happened in there as well as the geographic regions of higher incidence that might "infect" their neighbors.

Since Log-Gaussian Cox processes simulated geographic regions of higher incidence, temporal abstracts might be useful. Hence, we chose the **Temporal Frequency Rate** that measures for each spatial granule the percentage of atoms occurred on it given all the atoms of the phenomenon at a given LoD.

In order to capture the $LoDs_{st}$ where the Log-Gaussian Cox process is better perceived, we correlate the Compact Temporal Abstract - Coefficient of variation and the Spatial autocorrelation of the Temporal frequency rate as can be seen in Figure 6.18. These two Compact Temporal Abstracts are chosen because, we want to capture the $LoDs_{st}$ in which there is a considerable variation on the Temporal frequency rate, and simultaneously, to understand whether the spatial granules are spatially correlated on the Temporal frequency rate.

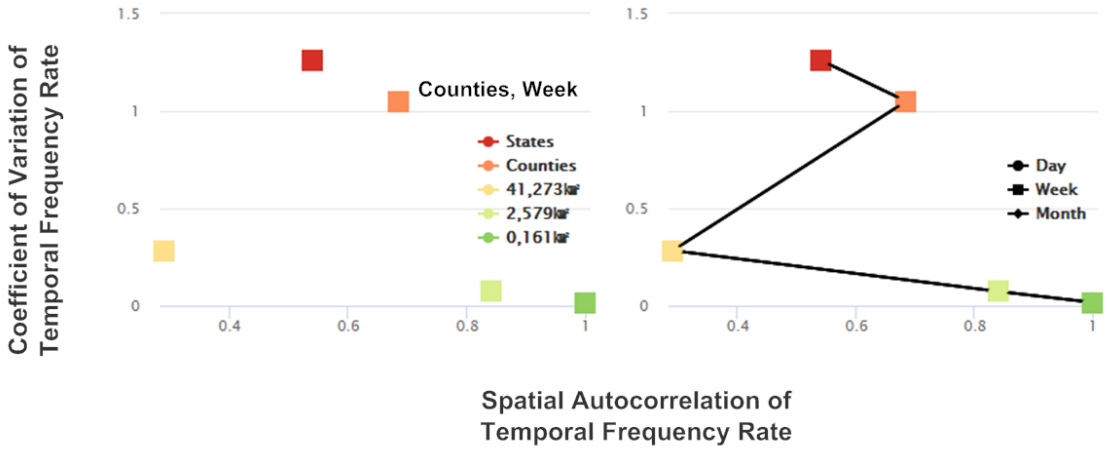


Figure 6.18: Dataset 7 (Log-Gaussian Cox process) - Correlation between the Coefficient of Variation of Temporal Frequency Rate and the Spatial Autocorrelation of Temporal Frequency Rate.

First of all, the temporal granularity does not have an impact on the **Temporal Frequency Rate**. Regardless the temporal granularity, the percentage of atoms occurred on particular spatial granules remains the same as can be observed on the left chart of Figure 6.18.

That being said, let's focus on the right chart in Figure 6.18. In finer spatial granularities, a spatial autocorrelation among spatial granules it is expected to exist, once their values diverge little or nothing as shown by their coefficient of variation. But when we look at the $LoDs_{st}$ - $(Counties, Week)$, the coefficient of variation is a value near to one, which indicates variability among values, and simultaneously, the level of spatial autocorrelation grows. But if we move to $LoDs_{st}$ - $(States, Week)$, the spatial autocorrelation decreases.

To check the previous analysis, the **Temporal Frequency Rate** is shown in Figure 6.19 at the $LoDs_{st}$ - $(Counties, Week)$.

There are some counties (that are spatially small) on the east side of USA (highlighted with a red arrow) that have a greater incidence of events. Such geographic area was zoomed-in and displayed at two $LoDs_{st}$: (i) $(Counties, Week)$; (ii) $(Raster\left(41.27km^2\right), Week)$ as shown in Figure 6.20.

Looking at the $LoDs_{st}$ - $(Raster\left(41.27km^2\right), Week)$ geographic regions with higher

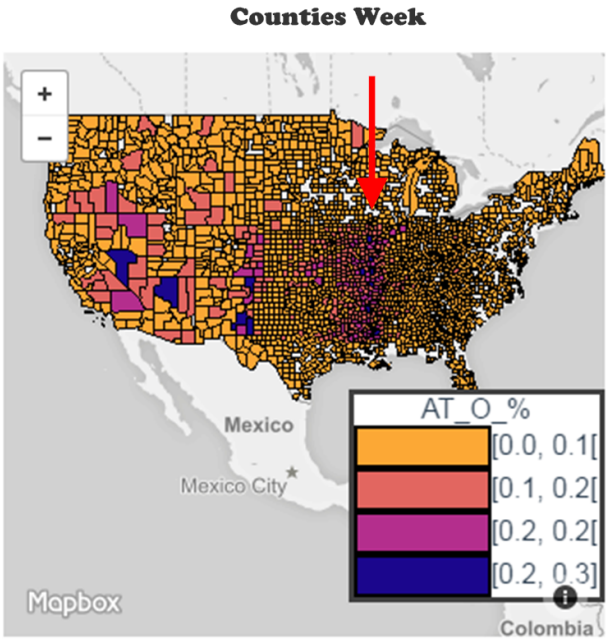


Figure 6.19: The **Temporal Frequency Rate** at the $LoDs_{st}$ - $(Counties, Week)$

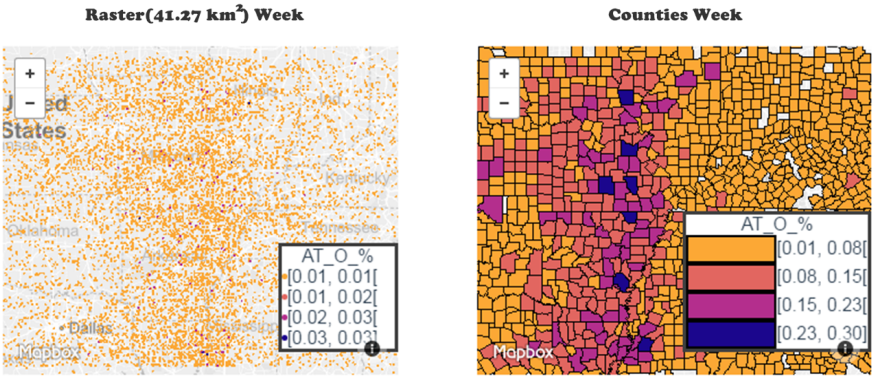


Figure 6.20: The **Temporal Frequency Rate** at the LoD_{st} - $(Raster\left(41.27km^2\right), Week)$ and $(Counties, Week)$.

incidence of events are no longer perceived. Although there are geographic regions with higher incidence (purple and dark blue spatial granules), the values of **Temporal Frequency Rate** are not different as they are in the $LoDs_{st}$ - $(Counties, Week)$. This confirms that the $LoDs_{st}$ - $(Counties, Week)$ is probably one of the suitable $LoDs_{st}$ to better understand the geographic regions that are most affected by the phenomenon.

6.3 Results on Real Datasets

Several phenomena were analyzed using the SUITE-VA. As opposed to synthetic datasets, we are not aware of possible patterns that those phenomena might contain. The phenomena collected were: (*i*) forest fires in Portugal; (*ii*) the dataset made public by the Armed

Conflict Location and Event Data Project[2] about conflict and protest data, occurring in Africa; (*iv*) robberies in the city of Chicago; (*iii*). These datasets contain information about different phenomena occurring in different spatial extents and different temporal extents.

6.3.1 Forest Fires in Portugal

This section shows the analysis made about wildfires that occurred in Portugal between 2001 and 2012. The granularities-based model was used in order to model them at different LoDs. This phenomenon is described by a collection of 280.968 spatiotemporal events. These events were modeled through the wildfires predicate containing two arguments $wildfires(space, time)$.

The most detailed spatial granularity $Raster\left(0.005km^2\right)$ is based on grid of 16384x 16384 cells that cover the analyzed spatial extent of the phenomenon, and each cell has an area of 0.005 km^2.

The remaining raster granularities for space were granularities with cell sizes of 0.08 km^2 and 0.319 km^2. The granularities *Parishes*, *Counties* and *Districts* was also considered. The time granularities used were *Hours*, *Days*, *Weeks*, *Months*, *Years*.

The considered granular terms required to model these events were: *Instant* and *Cell* for the time and space arguments, correspondingly. The raw data were encoded at the base LoD of the wildfires predicate which includes the time granularity of *Hours* and the space granularity $Raster\left(0.005km^2\right)$.

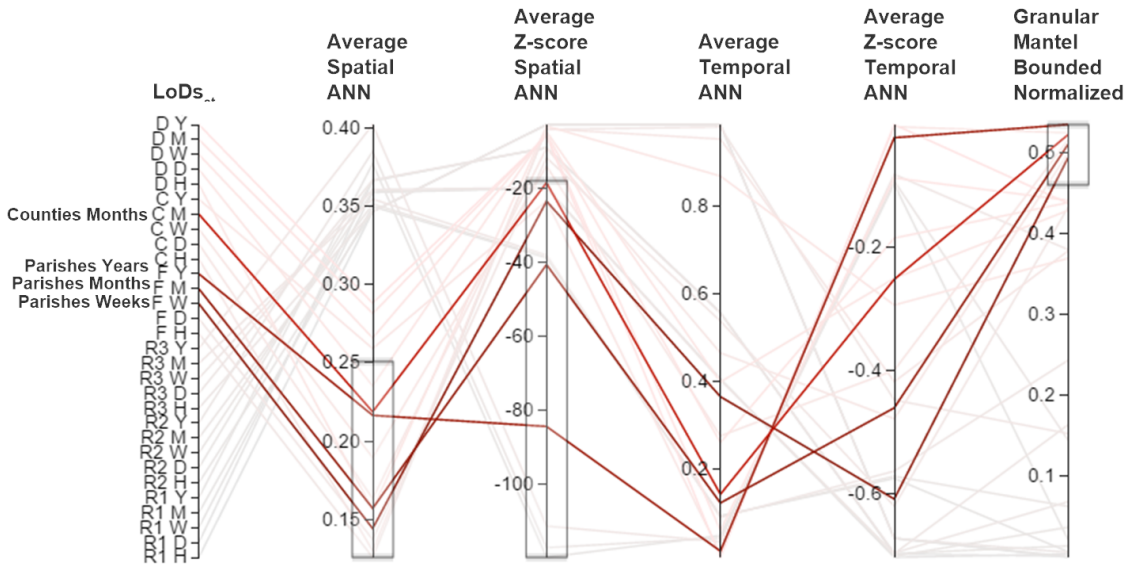


Figure 6.21: Overview of wildfires in Portugal using Global Abstracts, Compact Spatial Abstracts and Compact Temporal Abstracts.

In order to have a grasp of wildfires in Portugal, we chose the following abstracts: (*i*) the GMBN; (*ii*) the average of Spatial ANN; (*iii*) the average of the z-score of the Spatial

[2]Website: `http://www.acleddata.com/`

ANN; (*iv*) the average of Temporal ANN; (*v*) the average of the z-score of the Temporal ANN. The Parallel Coordinates was used to simultaneously analyze the global abstracts chosen across all the $LoDs_{st}$. Let's take a close look at them:

There are $LoDs_{st}$ holding values close to zero with respect to Average Spatial ANN that simultaneously have quite negative values considering its z-score. Therefore, this kind of values have some resemblances with the ones obtained with Poisson cluster simulated datasets or the Contagious ones. As a result, at this point, we might say that wildfires in Portugal hardly follow a homogenous model.

Several $LoDs_{st}$ are holding values close to zero with respect to Average Temporal ANN but their z-scores are close to zero, which means that the complete randomness cannot be rejected. In other words, wildfires occurring on the same spatial granule are likely not close to each other in time, on average. Furthermore, this information is telling us that probably, we are not dealing with a phenomenon that follows a Contagious process.

Several $LoDs_{st}$ have the spatiotemporal interaction among events measured by the GMBN above 0.4, which is similar to the values obtained in Poisson Cluster simulated datasets. This reinforces the similarities of the wildfires in Portugal with the Poisson Cluster model.

Based on the preliminary analysis, wildfires in Portugal seem to approach the Poisson Cluster model. The Parallel Coordinates was filtered in order to identify the suitable $LoDs_{st}$ to confirm the previous hypothesis. We just considered $LoDs_{st}$ with values below 0.25 (approximately) regarding the Average of the Spatial ANN. For the average of its z-score, we just considered values below -20 (approximately). Finally, the top four values of the GMBN were considered, which means values above 0.45 (approximately). The other coordinates (temporal average nearest neighbor and its z-score) were not filtered because there are no domain values that clearly points to clustered or dispersed events in time. From the filtering conducted, four $LoDs_{st}$ were highlighted: $(Parishes, Weeks)$, $(Parishes, Months)$, $(Parishes, Years)$, $(Counties, Months)$.

To better understand how wildfires, occur in space over time, the Spatial ANN and its z-score were plotted in a scatter plot for the $LoDs_{st}$ identified previously as can be seen in Figure 6.22. First of all, notice that, the charts obtained present similarities in the values and corresponding "shapes" with the charts obtained when we studied Poisson Cluster simulated datasets. Furthermore, in all $LoDs_{st}$ displayed, the phenomenon reveals to have several clustered distributions of events over time.

Nevertheless, $LoDs_{st}$ - $(Parishes, Weeks)$ (chart on the bottom-right) is the one that better fits the Poisson Cluster process/model. That is, in general, events occur near one another but there are a few times when events did not occur or occur in a dispersed way. Furthermore, in the $LoDs_{st}$ - $(Parishes, Weeks)$ there is a good tradeoff between

the Spatial ANN and its z-score. In other words, there are many temporal granules in which the Spatial ANN's values are around 0.15 (trend toward clustering) and where their z-scores are quite negative (confirmation of clustering).

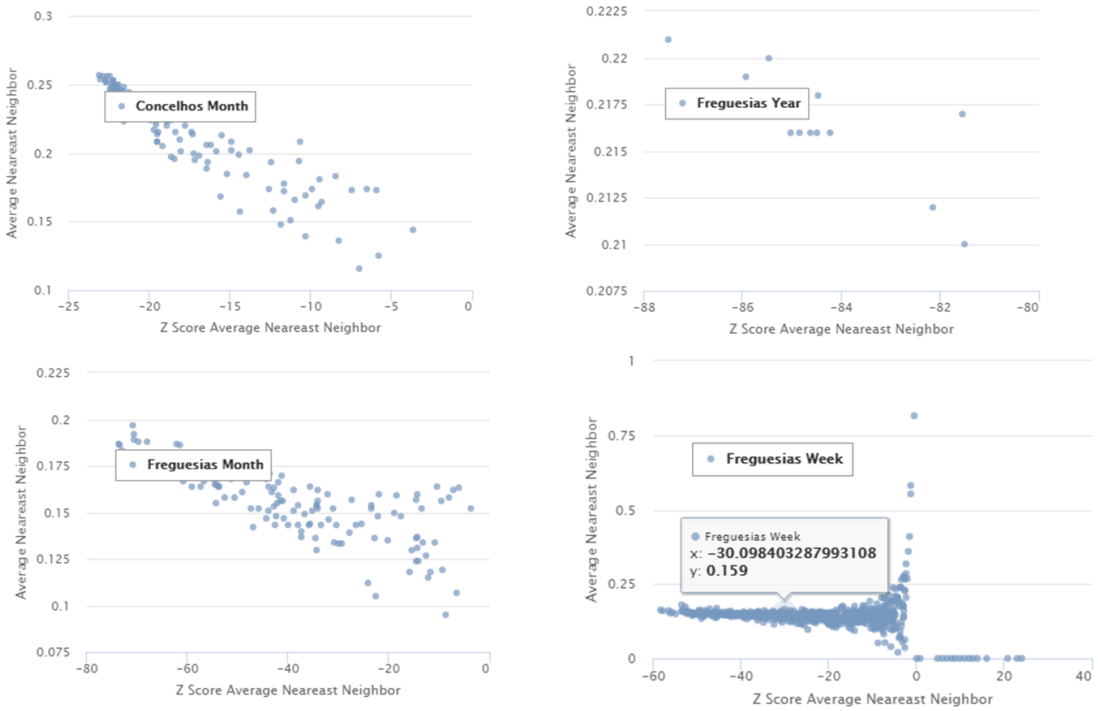


Figure 6.22: The Spatial ANN and its Z-score displayed in four $LoDs_{st}$.

The SUITE-VA allows users to zoom-in on a particular area of a scatter plot. When that action is performed the selected points (i.e., temporal granules) are highlighted on the corresponding time-series using vertical red lines. Thus, to understand when wildfires are occurring spatially clustered, we zoom-in the scatter plot at $LoDs_{st}$ - $(Parishes, Weeks)$ over the area where the Spatial ANN is less than 0.2 and its z-score is less than -35. So, we are choosing the temporal granules in which the events are more spatially clustered. The result of this selection can be seen Figure 6.23.

The time series on the right-hand side is showing the entire temporal period at study. From it, we can notice that the wildfires occurred recurrently spatially clustered which, in general, matches the summer periods but not necessarily. For instance, during the week that has started on November, 5th 2011, several wildfires occurred in Portugal. These are displayed on the map of Figure 6.23, and it is possible to confirm that they are spatially clustered.

6.3.2 Violence against Civilians in Africa

This section shows the analysis made over violence against civilians in Africa that occurred between 1997 and 2015. The granularities-based model was used in order to model them at different LoDs. This phenomenon is described by a collection of 33.393

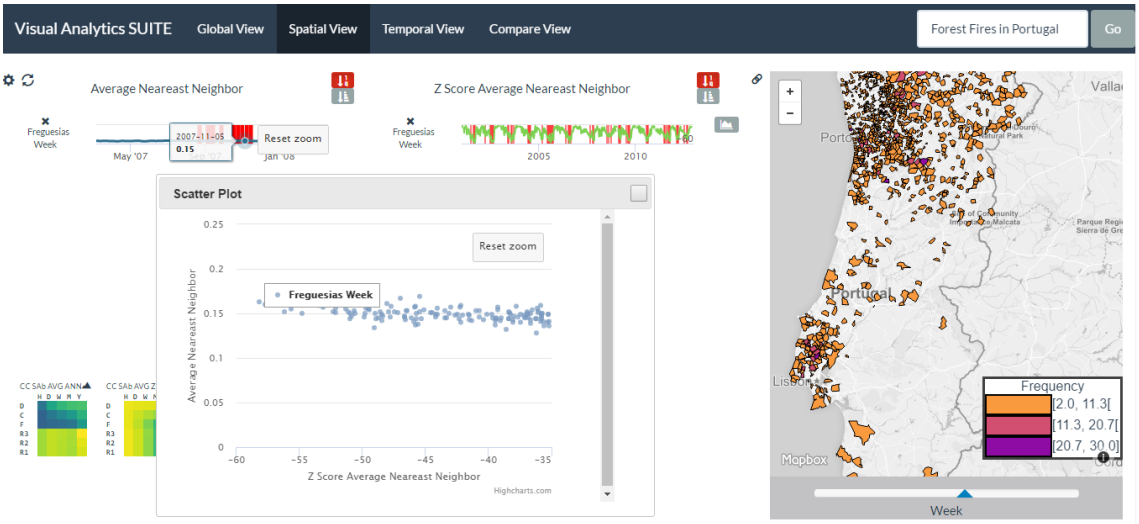


Figure 6.23: Filter the temporal granules in which the clusters of events are most pronounced at $LoDs_{st}$ - $(Parishes, Weeks)$.

spatiotemporal events. These events were modeled through a terrorism predicate, with two arguments $terrorism(space, time)$.

The most detailed spatial granularity $Raster\left(343.45km^2\right)$ is based on a grid of 16384 x 16384 cells that cover the analyzed spatial extent of the phenomenon, and each cell has an area of 343.45 km^2. The other coarser spatial granularities were obtained by dividing the number of cells in the grid by a factor of 2. So the valid granularities for space were rasters with cell sizes of 1376.34 km^2, 5525.79 km^2, and 22268.15 km^2. The used time granularities were $Hours$, $Days$, $Weeks$, $Months$, $Years$.

Like previously, we start by trying to figure out what kind of model might be underlying this phenomenon using the usual abstracts: (i) the GMBN; (ii) the Average of Spatial ANN; (iii) the Average of the z-score of the Spatial ANN; (iv) the Average of Temporal ANN; (v) the Average of the z-score of the Temporal ANN. The Parallel Coordinates are displayed in Figure 6.24.

There are "four levels" of spatial clustering over time as depicted by the Average Spatial ANN and the corresponding z-scores. These levels are being strongly influenced by the temporal granularity. With the granularity $Years$, the Average Spatial ANN and the corresponding z-scores reach their minimums while with the granularity $Days$ the spatial clustering over time is not so pronounced. Thus, the phenomenon seems to have some similarities to the Poisson Cluster model.

Several $LoDs_{st}$ are holding values close to zero with respect to Average Temporal ANN but their z-scores are also close to zero, which means that the complete randomness cannot be rejected. In other words, the attacks against civilians occurring on the same spatial granule are likely not close to each other in time, on average. Furthermore, this information is telling us likely, we are not dealing with a phenomenon

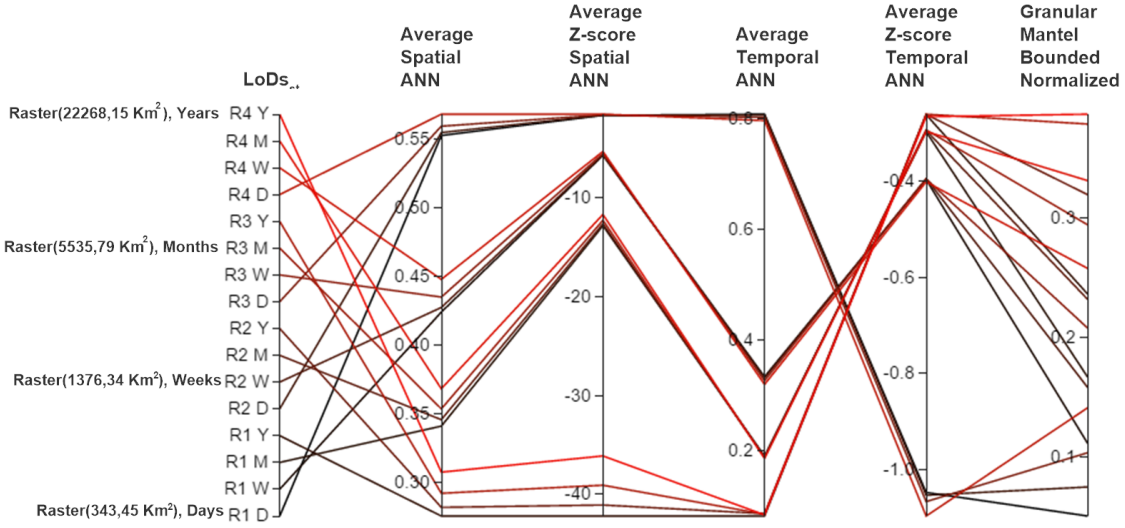


Figure 6.24: Overview of the attacks against civilians in Africa using Global Abstracts, Compact Spatial Abstracts and Compact Temporal Abstracts.

that follows a Contagious process. This is quite similar to the phenomenon about Wildfires in Portugal.

There are some $LoDs_{st}$ that have the spatiotemporal interaction among events measured by the GMBN above 0.3, which is similar to the datasets simulated with the Poisson cluster or with the dataset about wildfires in Portugal.

Since the Average Spatial ANN and the corresponding z-scores reach their minimums, we look to the data at the $LoDs_{st}$ - $(Raster(22268.15\,km^2), Years)$ in three temporal granules: 2008, 2009, 2010. The temporal granules were chosen for no particular reason but just to see if there were clusters of events based on the tip provided by the Parallel Coordinates (see Figure 6.24).

As we can see in Figure 6.25, the spatiotemporal events are in fact spatially clustered. In this case, there are clusters of events that remain stable in the three years chosen like for example the cluster in Mozambique (green circle), South Nigeria (red circle), and on the border between Uganda and Kenya (blue circle).

In our initial analysis about violence against civilians, the $LoDs_{st}$ containing the temporal granularity $Months$ also suggest some characteristics of the Poisson Cluster process, and consequently, clusters of events over time. So, we have chosen the $LoDs_{st}$ - $(Raster(22268.15\,km^2), Months)$ for displaying the Spatial ANN and the corresponding z-score. Afterward, we plot them in a scatter plot and filter out the temporal granules where the values of Spatial ANN are low and the values of the z-score are more negative, that is, the temporal granules where the clusters are likely most pronounced. This action highlights the time series on the respective granules as displayed in Figure 6.26.

Surprisingly, only "recent" temporal granules were highlighted which means that the attacks against civilians in Africa are getting more spatially clustered than in

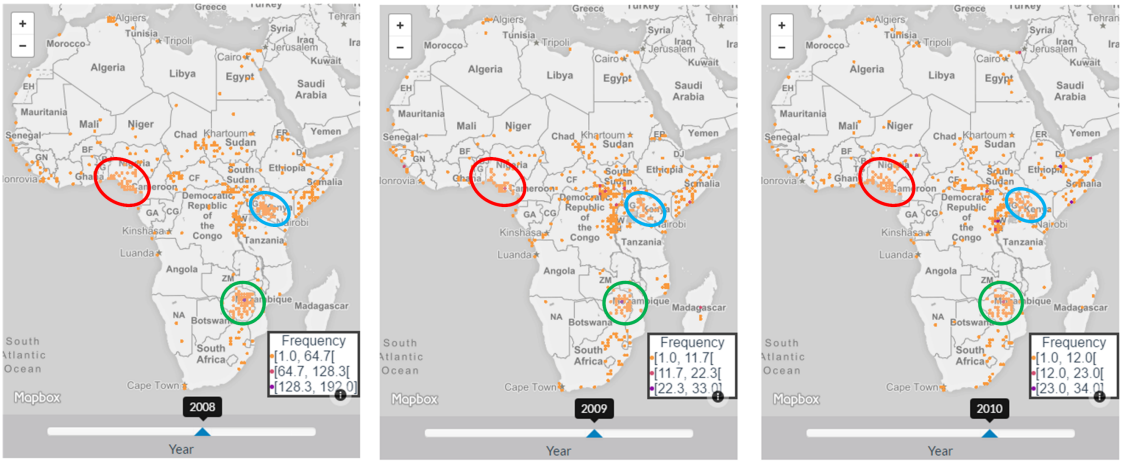


Figure 6.25: Violence against Civilians at the $LoDs_{st}$ - $Raster(22268.15\ km^2)$, $Years$ displayed in three temporal granules - 2008,2009 and 2010.

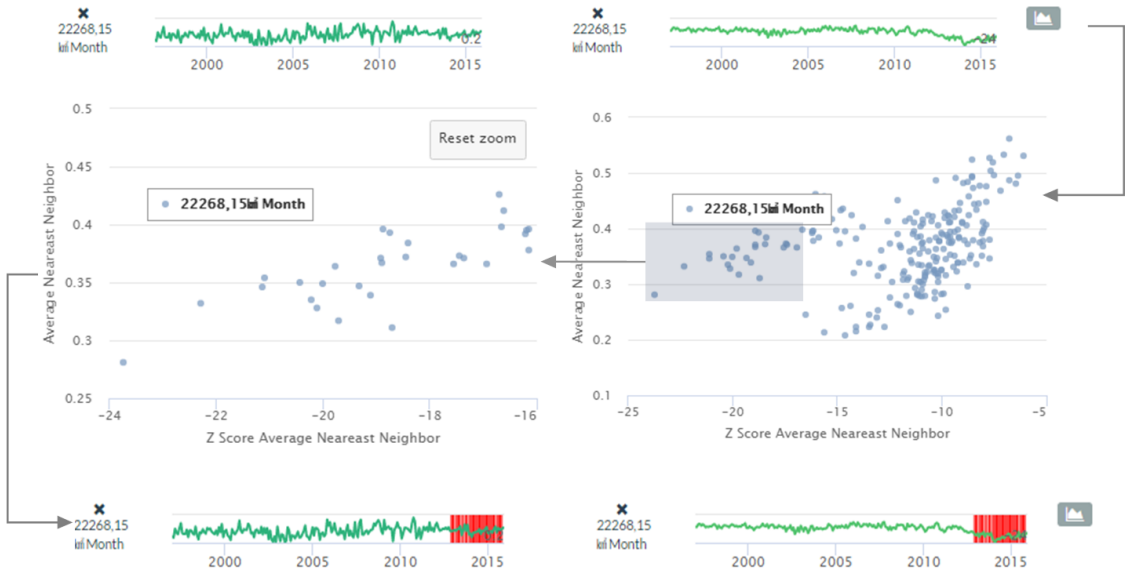


Figure 6.26: Highlighting the temporal granules where the Violence against Civilians is more spatially clustered using the $LoDs_{st}$ - ($Raster(22268.15\ km^2)$, $Months$).

the past. A more detailed analysis about this change can be done and the $LoDs_{st}$ - ($Raster(22268.15\ km^2)$, $Months$) seems suitable for such an analytical task.

Another experiment was made to understand whether the attacks against civilians occur on the same regions over time or if there were changes. To conduct this analysis, we have chosen the Temporal Center Mass's Positioning. Since the spatial autocorrelation of the Temporal Center Mass's Positioning is most pronounced in $LoDs_{st}$ containing the spatial granularity $Raster(343.45\ km^2)$, our analysis was detailed in the $LoDs_{st}$ - $Raster(343.45\ km^2)$, $Weeks$. Some results were displayed in Figure 6.27.

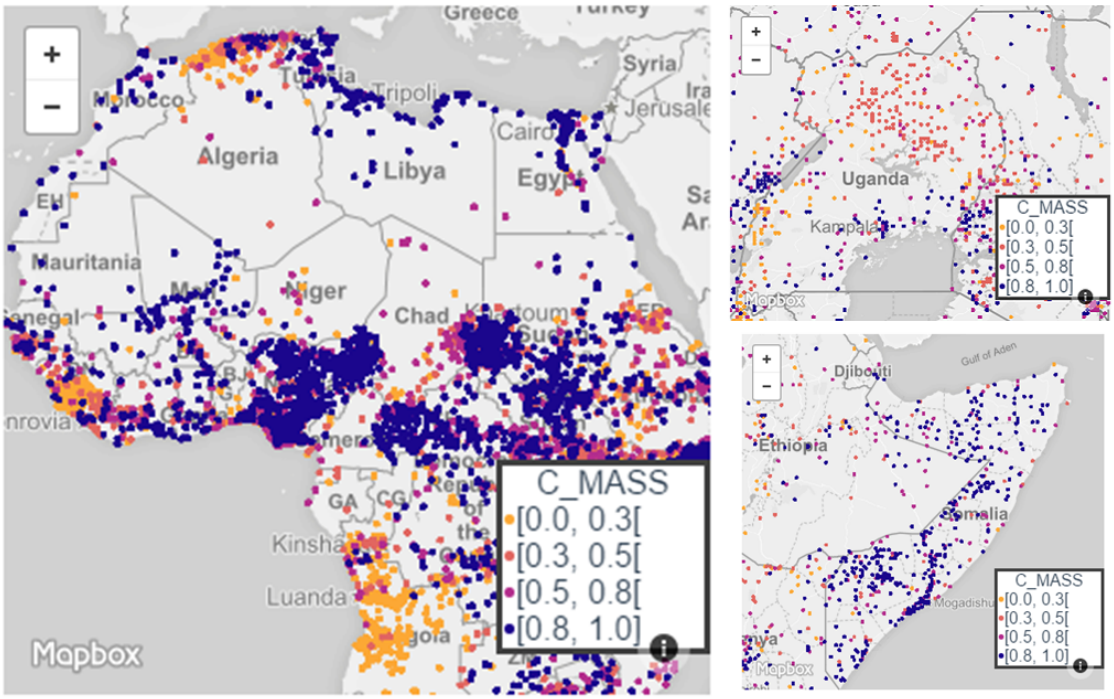


Figure 6.27: Violence against Civilians at the $LoDs_{st}$ - $Raster(343.45\ km^2)$, $Weeks$ displayed in three different spatial extents.

Clusters of events are changing over time. For instance, in Angola most of the attacks occurred in the past and they are not that frequent anymore. The same is observed in Serra Leoa. But for instance, at northern Algeria, the attacks have slightly changed location over time from north-west to north-east. Looking at the north of Uganda (top right-hand map), there is no particular pattern, that is, in each spatial granule there may be old and recently attacks or the attacks happened somewhere in the middle of the period under study (1997-2015). At Somalia, most attacks are recent and spread out by the entire country.

In the case of Nigeria, we have used a feature of the SUITE-VA, which allows to hide and show the events holding a particular class of values. In Figure 6.28, from left to right, the classes were incrementally added to the map. As you can see, in the past, most attacks

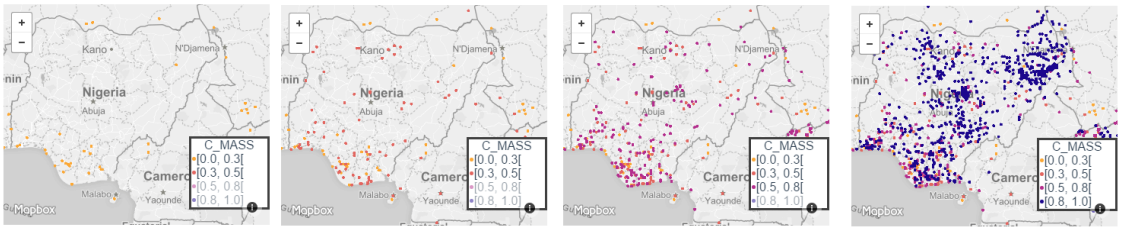


Figure 6.28: Evolution of Violence against Civilians throughout time at the $LoDs_{st}$ - *Raster*$(343.45\ km^2)$, *Weeks* in Nigeria.

occurred at the south of Nigeria, and afterward, they started to spread across the entire country.

6.3.3 Robberies in Chicago

This section shows the analysis made over robberies that happened in the City of Chicago and occurred between 2001 and 2015. The granularities-based model was used in order to model them at different LoDs. This phenomenon is described by a collection of 221.625 spatiotemporal events. These events were modeled through a robberies predicate, with two arguments $robberies(space, time)$.

The most detailed spatial granularity $Raster\left(0.002km^2\right)$ is based on grid of 1024x 1024 cells that cover the analyzed spatial extent of the phenomenon, and each cell has an area of 0.002 km^2. The other coarser spatial granularities were obtained by dividing by a factor of 2 the number of cells in the grid. So the valid granularities for space were rasters with cell sizes of 0.007 km^2, 0.027 km^2. The granularity *Chicago's Community areas* was also considered. The time granularities used were *Hours*, *Days*, *Weeks*, *Months*, *Years*.

The considered granular terms required to model these events were: *Instant* and *Cell* for the time and space arguments, correspondingly. The raw data were encoded at the base LoD of the robberies predicate which includes the time granularity of *Hours* and the space granularity $Raster\left(0.007km^2\right)$.

Like previously, we start by trying to figure out what kind of model might be underlying this phenomenon using the following abstracts: (*i*) the GMBN; (*ii*) the Average of Spatial ANN; (*iii*) the Average of the z-score of the Spatial ANN; (*iv*) the Average of Temporal ANN; (*v*) the Average of the z-score of the Temporal ANN.

The Parallel Coordinates are displayed in Figure 6.29. Based on it, we found the following:

In what concerns the Average of Spatial ANN, the LoD_{st} - (*Raster*$(0.002\ km^2)$, *Months*) (*R1 M* in Figure) holds the minimum value that corresponds to 0.51 (and its z-score is -31.84). All the others $LoDs_{st}$ have a greater value of the Average of Spatial ANN, and therefore, it seems that robberies in City of Chicago do not follow the Poisson Cluster process.

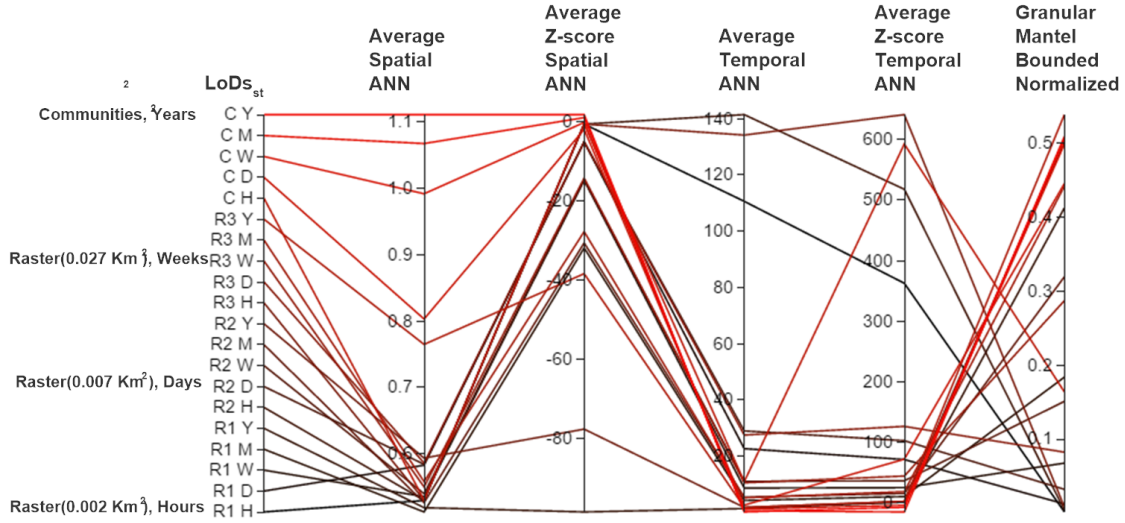


Figure 6.29: Overview about robberies in the City of Chicago using Global Abstracts, Compact Spatial Abstracts and Compact Temporal Abstracts.

All the $LoDs_{st}$ are holding values above one with respect to Average Temporal ANN. In other words, robberies occurring on the same spatial granule are likely not close to each other in time, on average. Therefore, this phenomenon hardly follows a Contagious process.

Looking at the GMBN, the $LoDs_{st}$ with higher values (around 0.5) are the ones that have the granularity *Communities* (*C* in Figure). This is actually similar to the Log-Gaussian Cox process discussed in Section 6.2.3. That is, a considerable evidence of spatiotemporal interaction, and weak/no evidence of the Poisson Cluster and Contagious process.

The previous analyzes make us look at the Temporal Abstract - Temporal Frequency in order to understand if the phenomenon follows a Log-Gaussian Cox process, that is, if there are communities areas with a higher incidence of robberies as pointed out by the previous analysis. Notice that, there are 77 communities areas and if the robberies had been distributed evenly then in each community area 1.3% robberies would have happened in each community area (approximately).

Since the temporal granularity does not have an impact on the Temporal Frequency Rate (as explained previously), then the Temporal Abstract - Temporal Frequency will be the same in any LoD_{st} containing the granularity *Communities*. Figure 6.30 displays the Temporal Frequency at the LoD_{st} - (*Communities*, *Months*).

There are in fact some communities areas with higher incidence of robberies that are close to each other. Austin, Auburn Gresham and South Shore communities areas are the ones with a higher incidence of robberies. In Austin 7% of all robberies happened, in South Shore 3.9% and in Auburn Gresham 3.52%.

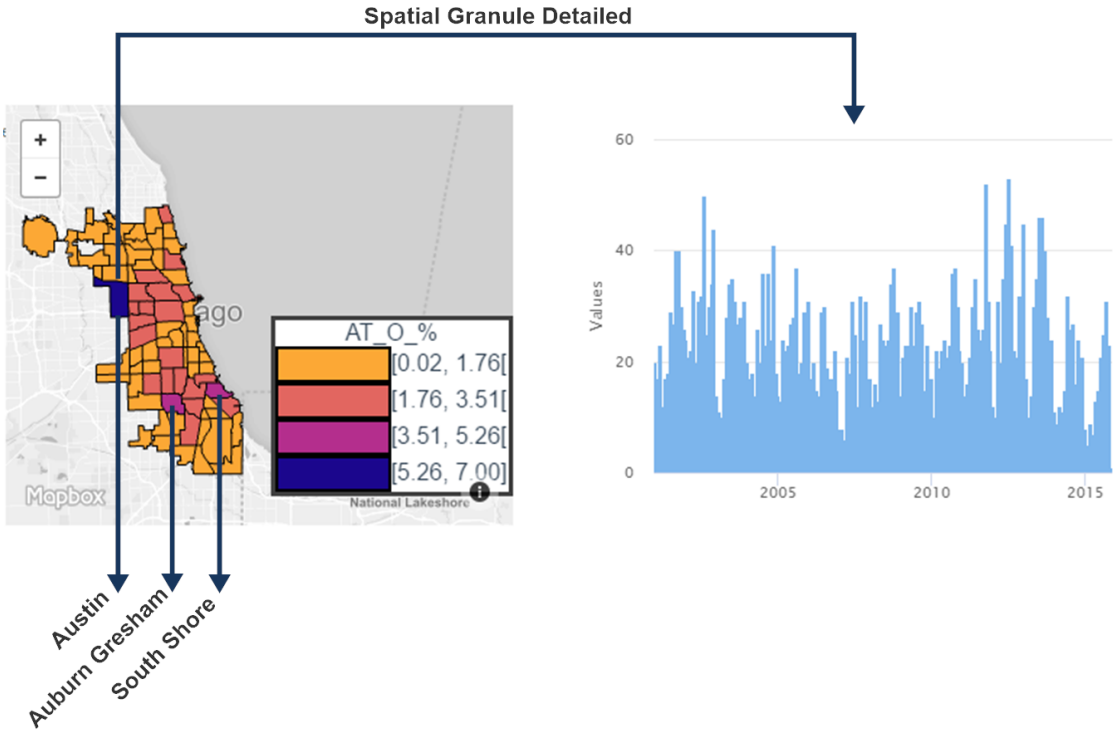


Figure 6.30: The Temporal Frequency Rate at the LoD_{st} - (*Communites*, *Month*).

Using the SUITE-VA, a user can detail a particular value of a Temporal Abstract by clicking on the spatial granule that we want to detail (in the map). Such action will show in the Phenomenon Representation area an interactive histogram that displays the distribution of the number of events by temporal granules concerning only the spatial granule selected.

In Figure 6.30, the spatial granule Austin was detailed. Interacting with the histogram, we concluded that, in Austin, the months of October, September, December and January are the months when the pikes normally happen.

CHAPTER 7

Conclusions and Future Work

This chapter closes the document. Section 7.1 summarizes the main contributions presented in the previous chapters while Section 7.2 discusses possible future work directions.

7.1 Conclusions

Many phenomena are being logged as spatiotemporal events at high levels of detail (LoDs), allowing us to better understand natural phenomena or human activities occurring on the surface of the Earth. Present-day visual analytics (VA) approaches targeting the analysis of spatiotemporal events present two main issues.

In general, VA approaches support separate analyzes of the spatial and temporal dimensions of the events that can prevent us from discovering information in datasets of spatiotemporal events. A typical form of analysis in such conditions is counting the number of events per day or per month. But a lot of information about the spatiotemporal dynamics of events arises when one works with the spatial and temporal dimensions together. An example of such an analysis is the computation of the spatiotemporal interaction level among spatiotemporal events.

Some of the VA approaches support analyzes that search for spatiotemporal patterns. But such approaches are commonly developed for a particular application domain that looks for a specific spatiotemporal pattern. In a way, those approaches assume the preexistence of the pattern thus probably leaving aside other patterns that might be discovered. Furthermore, VA approaches follow an analysis approach based on a single LoD, and therefore, the users have to choose the appropriate LoDs to perform the analysis of the data. Thus, when there is little information about a spatiotemporal phenomenon, i.e., an early stage of analysis, the user does not only ignore what patterns might be on the

data but also what is/are the suitable LoD(s) to find them. This led us to formulate the following research question:

> *How can we help users explore phenomena logged as spatiotemporal events across multiple LoDs, simultaneously, helping them to understand in what LoDs there are patterns emerging?*

To enhance the analyses over spatiotemporal events, we first propose to move from a single user-driven LoD to a multiple LoDs analysis approach, providing the user with an understandable high-level overview of the underlying structure of the phenomenon for each LoD. This approach can provide several hints about the different facets of spatiotemporal events that can provide a first insight on the presence or absence of patterns at particular LoDs. Following this approach, we aimed to help users to detect very soon in what LoDs there are potential patterns and what kind of patterns they are. According to his analytical goal and domain knowledge, the user may be able to better guide his analysis thus avoiding an information overload.

A long path was made to support analyses at multiple LoDs, simultaneously, over spatiotemporal events. A natural requirement for that is the capability to represent and reason about spatiotemporal events at different LoDs. To meet this need, this PhD thesis contributes with a **Theory of Granularities (ToG)** that supports granularities defined over any data domain covering the definitions proposed in the literature. The ToG proposed introduces four induced relations in order to transpose the relations defined in the domains of reference for granules belonging to granularities. None of the works discussed in the literature can do that, as detailed in Section 3.

Although the ToG supports the creation of granularities over any data domain that can be used to describe spatiotemporal events, we needed to model phenomena logged as spatiotemporal events at several LoDs. To meet this need, a granular computing approach was proposed to model spatiotemporal phenomena at multiple LoDs, labeled as **the granularities-based model**.

The granularities-based model lies on the concept of LoD which is a key contribution of this work. The granularities-based model follows an automated approach to generalize a phenomenon from one LoD to a coarser one. When changing a phenomenon's LoD a time interval can eventually be generalized to a time instant while a region might be simplified. To the best of our knowledge, there is no other model like the one proposed here as detailed in Section 4.5.

Granular terms are used in statements' arguments that allow expressing abstract real world entities in a granular way. Based on the general concept of granular term, spatial granular terms (Cell and RasterRegion) and temporal granular terms (Instant and Interval) were formalized. Regarding the latter, we transpose the temporal topological relations to the temporal granular terms. A theoretical analysis was made for reasoning about what happens to the topological relations (in the context of temporal granular

terms) when these are generalized. This work ends up being a complementary contribution because it extends a previous work as detailed in Appendix B.

Through the granularities-based model, there is a phenomenon's representation for each LoD. To provide a theoretical foundation that anchors analyses at multiple LoDs, simultaneously, **this PhD Thesis contributes with a framework for SUmmarizIng spatioTemporal Events (SUITE)**. To the best of our knowledge, there are no approaches that work across several spatial and temporal LoDs and that are independent of the application domain in the context of spatiotemporal events as discussed in Section 5.6.

SUITE was developed on top of the granularities-based model and builds summaries, at different LoDs, about phenomena logged as spatiotemporal events. The framework establishes five types of summaries working with space and time together. This allows us to frame and extend many proposals in the literature that create summaries of data in the proposed framework. But also, it allows that new summaries are proposed. In this work, we propose several new summaries that are detailed in Section 5.5. In particular, we propose the Granular Mantel Bounded and Normalized in order to handle the difficulty of existing spatiotemporal interaction methods on providing comparable values among LoDs as discussed in Appendix D.

To conduct analyzes in this new mindset, a web-based VA tool anchored on the SUITE framework was developed, designated by **SUITE-VA**. The tool allows to visually inspect hints about the absence or presence of different kinds of spatiotemporal patterns at multiple LoDs, simultaneously, following a coordinated strategy among the visualizations provided. To the best of our knowledge, there is no other prototype or application that supports analyses over spatiotemporal events at multiple LoDs, simultaneously, following the VA Mantra as detailed in Section 6.1.

The evaluation of our proposals was conducted with two types of datasets of spatiotemporal events: (i) synthetic datasets; (ii) real datasets. Synthetic datasets with different spatiotemporal patterns (i.e., homogenous process, Poisson-cluster process, Contagious process, Log-gaussian Cox process) in different spatiotemporal LoDs, with different cardinality were produced. For most of the datasets produced, the SUITE-VA could provide a correct overview of the "phenomenon" allowing us to identify the LoD(s) in which the pattern generated occurs, and therefore, the LoDs that should be used to detail the analysis as detailed in Section 6.2.

We then look for the patterns/processes identified previously in real datasets. The real datasets studied were: (i) forest fires in Portugal; (ii) violent attacks against civilians occurring in Africa; (iii) and, robberies in the city of Chicago. Recognizing some of the processes in these datasets, in different spatiotemporal LoDs, was easy.

Forest fires in Portugal have similarities with a Poisson-cluster model process, which is better perceived at the spatiotemporal LoD $Parishes, Weeks$. Nevertheless, there is a difference between the periods of time where the events happened in a dispersed form and in a clustered way. In general, wildfires occurred in a clustered form during summer seasons while they happened mainly in a dispersed form in the other seasons. Despite

that, some temporal outliers were found like for example the week that has started on November, 5th 2011 in which several wildfires occurred mainly in the north of Portugal (see Section 6.3).

Similarly, violent attacks against civilians in Africa have similarities with a Poisson-cluster process. In this case, the analysis suggests that the pattern is better perceived in the spatiotemporal LoDs containing the temporal granularities *Months* or *Years*. In such spatiotemporal LoDs more analyses were conducted and some findings are reported. For example, in Angola most of the attacks occurred in the past and they are not very frequent anymore. The same is observed in Serra Leone. But for instance, in north Algeria, the location of the attacks has changed slightly over time from northwest to northeast. Furthermore, a peculiar change has been detected in the phenomenon. It seems that, globally, the attacks against civilians in Africa are getting more spatially clustered than in the past as detailed in Section 6.3. This seems a relevant change that deserves a more detailed analysis.

Finally, robberies in the City of Chicago have similarities with a Log-Gaussian Cox Process, which is better perceived in the spatiotemporal LoDs containing the spatial granularity *Communities Areas*. Then, the spatiotemporal LoD *Communities Areas*, *Months* was detailed. The community areas in Chicago which are more prone to robberies are Austin, South Shore and Auburn Gresham.

From the experiments conducted and the results achieved, the SUITE-VA was able not only to provide an overview of the presence or absence of different spatiotemporal patterns but also suggest appropriate spatiotemporal LoDs that allow us to better perceive the corresponding patterns. That being said, it is reasonable to state that this PhD thesis enhanced the exploratory analysis of spatiotemporal events across multiple LoDs.

In Introduction 1.2 some questions were formulated that needed to be answered to address the research problem identified in this PhD thesis.

1. How do we enable representation and reasoning about spatiotemporal events at different LoDs? **Using the Theory of Granularities.**

   Making analyses across multiple LoDs requires modeling spatiotemporal events at different LoDs.

   a) What is a LoD? How do we formalize the concept of LoD? **A LoD is a set of argument pairs and a valid granularity with respect to a predicate *P*.**

   b) How do we model a phenomenon at different LoDs?

      Using the granularities-based model.

   c) Datasets of spatiotemporal events are collected at high LoDs. How do we follow a bottom-up automated approach in order to provide different phenomena's representations for each LoD?

Each function symbol has its own generalization rules. Atoms' granular terms of the granularities-based model are automatically generalized based on those rules.

2. With the datasets of spatiotemporal events available at multiple LoDs, we aim to provide analyses across them.

   a) How do we provide an understandable high-level overview about the underlying structure of the phenomenon for each LoD?

   Through the Abstracts proposed in the SUITE framework.

   b) How will the users inspect and compare the phenomenon perception across multiple LoDs?

   Leveraging from the SUITE-VA which encodes the several abstracts implemented and anchored by the SUITE framework into visualizations.

   c) How do we provide an approach independent from the phenomenon without focusing on a particular analytical task or pattern?

   Our approach is independent of the application domain because the SUITE lies on a granular computing approach (i.e., granularities-based model) which is independent from the application domain.

   Our approach does not focus on a particular analytical task or pattern as the SUITE framework establishes five types of summaries working with space and time that can be used to implement a function that measures any facet or pattern of a phenomenon.

7.2 Future Work

Our research is focused on enhancing exploratory analysis of spatiotemporal events. To accomplish that, we introduce a novel mindset that is devised to give an overview of potential patterns that might be in the data, and simultaneously, tell what LoDs are suitable to study them. Underlying the novel approach, the Theory of Granularities (ToG), the granularities-based model, the SUITE framework, and the SUITE-VA were proposed. In a way, this PhD thesis is not about a novel algorithm to look for a spatiotemporal pattern, or a novel visualization method to improve the execution of a particular analysis, or even a new data structure to compress spatiotemporal events in-memory in order to improve performance. Hence, we envisage several research directions that can be pursued as future work. These directions can be divided into theoretical foundations and applicational.

In what concerns theoretical foundations, we envisaged three main topics to further research:

- The Theory of Granularities can be extended. In particular, the definition of granularities based on other granularities and the concept of evolution of granularity should be considered.

- More work about topological relations between granular terms. On one hand, one can study what happens to topological relations between spatial granular terms proposed. On the other hand, this kind of studies can be generalized to account for different granularities in which the granular terms are defined.

- Evaluate other approaches of generalization in the granularities-based model.

From the applicational point of view, heuristics to suggest automatically LoDs to analyze the data are needed and should be a priority because if the number of abstracts grows considerably it might be overwhelming to the user. This issue relates to the learning curve. Each abstract looks for a feature or pattern which frequently is expressed in terms of a range of values. According to the value, it means one thing or the other. Thus, a user needs to get familiar with the abstracts and their interpretation. Requiring a user to memorize all the abstracts and their interpretation might be overwhelming, specially if we consider the joint interpretation of abstracts. So again, heuristics to suggest automatically LoDs should be a priority.

Leading the automatism of the suggestion to the extreme, one might considers the usage of supervised learning algorithms on labeled datasets (i.e., pattern and LoDs) to train a model composed by N abstracts and M LoDs as a way of predicting the pattern and the LoD based on the abstracts' values.

Another research direction can be the development of new abstracts that measure different spatiotemporal patterns or facets.

The tool does not make any parallel computation, that is, either the computation responsible for the generalization of the phenomenon or the computation of abstracts occurs on a single machine. The performance was not a concern of this PhD thesis. But notice that, we are just dealing with the performance at pre-computation phase. This way, a possible research direction is to consider NoSQL databases and employ parallel computing techniques in order to avoid not only the pre-computation of the phenomenon's representation at a particular LoD but also the pre-computation of abstracts. Several advantages might come from this. Such approach would mean that a user would be able to filter the events by semantic attributes. Thus, a user would be able to identify different patterns according to attributes' values. Notice that, this is not completely solved in the literature. For instance, the work (Swedberg and Peuquet 2016) supports slices by semantic attributes but their approach has good performance only up to 20.000 events (Swedberg and Peuquet 2016). On the other hand, a user would be able to define granularities on demand.

To end, the analysis across LoDs introduced was employed on spatiotemporal events. A similar research can be conducted but now for other application domains that require other spatiotemporal datatypes, like trajectories for example.

Bibliography

Moran, P. A. P. (1950). "Notes on continuous stochastic phenomena". In: *Biometrika*, pp. 17–23.

Erikson, E. H. (1959). "Identity and the life cycle: Selected papers." In: *Psychological issues*.

Knox, E. G. and M. S. Bartlett (1964). "The detection of space-time interactions". In: *Applied Statistics*, pp. 25–30.

Mantel, N. (1967). "The detection of disease clustering and a generalized regression approach". In: *Cancer research* 27.2 Part 1, pp. 209–220.

Tobler, W. R. (1970). "A computer movie simulating urban growth in the Detroit region". In: *Economic geography* 46.sup1, pp. 234–240.

Vilain, M. B. (1982). "A System for Reasoning About Time." In: *AAAI*, pp. 197–201.

Allen, J. F. (1983). "Maintaining Knowledge About Temporal Intervals". In: *Commun. ACM* 26.11, pp. 832–843. ISSN: 0001-0782.

Atallah, M. J. (1983). "A linear time algorithm for the Hausdorff distance between convex polygons". In: *Information processing letters* 17.4, pp. 207–209.

Bertin, J. (1983). "Semiology of graphics: diagrams, networks, maps". In:

Allen, J. F. (1984). "Towards a general theory of action and time". In: *Artificial intelligence* 23.2, pp. 123–154.

Openshaw, S. and S Openshaw (1984). "The modifiable areal unit problem". In: Geo Abstracts University of East Anglia.

Ebdon, D. (1985). *Statistics in geography*.

Kong, T. Y. and A. Rosenfeld (1989). "Digital topology: Introduction and survey". In: *Computer Vision, Graphics, and Image Processing* 48.3, pp. 357–393.

Inselberg, A. and B. Dimsdale (1991). "Parallel coordinates". In: *Human-Machine Interactive Systems*. Springer, pp. 199–233.

Getis, A. (1992). "The Analysis of Spatial Association by Use of Distance Statistics". In: *Geographical Analysis* 24.3, pp. 189–206.

Egenhofer, M. J. and J. Sharma (1993). "Topological relations between regions in ρ^2 and $\mathbb{Z}^2$". In: *Advances in spatial databases*. Springer, pp. 316–336.

Jacquez, G. M. (1996). "A k nearest neighbour test for space–time interaction". In: *Statistics in medicine* 15.18, pp. 1935–1949.

Qi, Y. and J. Wu (1996). "Effects of changing spatial resolution on the results of landscape pattern analysis using spatial autocorrelation indices". In: *Landscape ecology* 11.1, pp. 39–49.

Frank, A. U. (1998a). "Different Types of Times in GIS". In: *Spatial and temporal reasoning*. Ed. by M. Egenhofer and R. Golledge. Oxford University Press, pp. 40–61.

Frank, A. U. (1998b). "Different types of "times" in GIS". In: *Spatial and temporal reasoning in geographic information systems*, pp. 40–62.

Stell, J. and M. Worboys (1998). "Stratified Map Spaces: A Formal Basis for Multi-resolution Spatial Databases". In: *Proceedings 8th International Symposium on Spatial Data Handling*. Department of Computer Science , Keele University , Staffordshire , UK ST5 5BG, pp. 180–189.

Zadeh, L. A. (1998). "Some reflections on soft computing, granular computing and their roles in the conception, design and utilization of information/intelligent systems". In: *Soft Computing-A fusion of foundations, methodologies and applications* 2.1, pp. 23–25.

Marceau, D. J. (1999). "The scale issue in the social and natural sciences". In: *Canadian Journal of Remote Sensing* 25.4, pp. 347–356.

Roddick, J. F. and M. Spiliopoulou (1999). "A bibliography of temporal, spatial and spatio-temporal data mining research". In: *ACM SIGKDD Explorations Newsletter* 1.1, pp. 34–38.

Weibel, R. and G. Dutton (1999). "Generalising spatial data and dealing with multiple representations". In: *Geographical information systems* 1, pp. 125–155.

Bettini, C, S Jajodia, and S Wang (2000). *Time Granularities in Databases, Data Mining, and Temporal Reasoning*. Springer. ISBN: 9783540669975.

Magnusson, M. S. (2000). "Discovering hidden time patterns in behavior: T-patterns and their detection". In: *Behavior Research Methods, Instruments, & Computers* 32.1, pp. 93–110.

Wu, J., D. E. Jelinski, M. Luck, and P. T. Tueller (2000). "Multiscale analysis of landscape heterogeneity: scale variance and pattern metrics". In: *Geographic Information Sciences* 6.1, pp. 6–19.

Chawla, S., S. Shekhar, W. Wu, and U. Ozesmi (2001). "Modeling Spatial Dependencies for Mining Geospatial Data." In: *SDM*. SIAM, pp. 1–17.

Erwig, M. and M. Schneider (2002). "Spatio-temporal predicates". In: *IEEE Transactions on Knowledge and Data Engineering* 14.4, pp. 881–901. ISSN: 1041-4347.

Tversky, B., J. B. Morrison, and M. Betrancourt (2002). "Animation: can it facilitate?" In: *International journal of human-computer studies* 57.4, pp. 247–262.

Bittner, T. and B. Smith (2003). "A theory of granular partitions". In: *Foundations of geographic information science* 7, pp. 124–125.

Harrower, M. and C. A. Brewer (2003). "ColorBrewer. org: an online tool for selecting colour schemes for maps". In: *The Cartographic Journal* 40.1, pp. 27–37.

Kraak, M.-J. and F. Ormeling (2003). "Cartography: visualisation of geospatial data". In: *Essex: Pearson Education Limited*.

Leipnik, M. R. and D. P. Albert (2003). *GIS in law enforcement: Implementation issues and case studies*. CRC Press.

Yao, X. (2003). "Research issues in spatio-temporal data mining". In: *Workshop on Geospatial Visualization and Knowledge Discovery, University Consortium for Geographic Information Science, Virginia*.

Andrienko, N. and G. Andrienko (2004). "Interactive visual tools to explore spatio-temporal variation". In: *Proceedings of the working conference on Advanced visual interfaces*. ACM, pp. 417–420.

Benz, U. C., P. Hofmann, G. Willhauck, I. Lingenfelder, and M. Heynen (2004). "Multi-resolution, object-oriented fuzzy analysis of remote sensing data for GIS-ready information". In: *ISPRS Journal of photogrammetry and remote sensing* 58.3, pp. 239–258.

Chen, H., W. Chung, J. J. Xu, G. Wang, Y. Qin, and M. Chau (2004). "Crime data mining: a general framework and some examples". In: *Computer* 37.4, pp. 50–56.

Fuchs, G. and H. Schumann (2004). "Visualizing abstract data on maps". In: *Information Visualisation, 2004. IV 2004. Proceedings. Eighth International Conference on*. IEEE, pp. 139–144.

Gatalsky, P., N. Andrienko, and G. Andrienko (2004). "Interactive analysis of event data using space-time cube". In: *Information Visualisation, 2004. IV 2004. Proceedings. Eighth International Conference on*. IEEE, pp. 145–152.

Keim, D. A., J. Schneidewind, and M. Sips (2004). "CircleView: a new approach for visualizing time-related multidimensional data sets". In: *Proceedings of the working conference on Advanced visual interfaces*. AVI '04. Gallipoli, Italy: ACM, pp. 179–182. ISBN: 1-58113-867-9.

Wang, S.-s. and D.-y. Liu (2004). "Spatio-temporal Database with Multi-granularities". In: *Advances in Web-Age Information Management SE - 15*. Ed. by Q. Li, G. Wang, and L. Feng. Vol. 3129. Lecture Notes in Computer Science. Springer Berlin Heidelberg, pp. 137–146. ISBN: 978-3-540-22418-1.

Worboys, M. F. and M. Duckham (2004). *GIS: a computing perspective*. CRC press.

Zhou, X., S. Prasher, S. Sun, and K. Xu (2004). "Multiresolution Spatial Databases: Making Web-Based Spatial Applications Faster". In: *Advanced Web Technologies and Applications SE - 5*. Ed. by J. Yu, X. Lin, H. Lu, and Y. Zhang. Vol. 3007. Lecture Notes in Computer Science. Springer Berlin Heidelberg, pp. 36–47. ISBN: 978-3-540-21371-0.

Dykes, J., A. MacEachren, and M. Kraak (2005). *Exploring geovisualization*. International Cartographic Association vol. 1. Elsevier. ISBN: 9780080445311.

Euzenat, J. and A. Montanari (2005). *Time granularity*.

Kapler, T. and W. Wright (2005). "Geotime information visualization". In: *Information Visualization* 4.2, pp. 136–146.

Keogh, E., J. Lin, and A. Fu (2005). "Hot sax: Efficiently finding the most unusual time series subsequence". In: *Data mining, fifth IEEE international conference on*. IEEE, 8–pp.

Ostfeld, R. S., G. E. Glass, and F. Keesing (2005). "Spatial epidemiology: an emerging (or re-emerging) discipline". In: *Trends in ecology & evolution* 20.6, pp. 328–336.

Ryden, K. (2005). "OpenGIS Implementation Specification for Geographic Information—Simple feature access—Part 1: common architecture". In: *OGC ref* 05-126.

Shanbhag, P., P. Rheingans, et al. (2005). "Temporal visualization of planning polygons for efficient partitioning of geo-spatial data". In: *Information Visualization, 2005. INFOVIS 2005. IEEE Symposium on*. IEEE, pp. 211–218.

Tominski, C., P. Schulze-Wollgast, and H. Schumann (2005). "3d information visualization for time dependent data on maps". In: *Information Visualisation, 2005. Proceedings. Ninth International Conference on*. IEEE, pp. 175–181.

Andrienko, N. and G. Andrienko (2006). *Exploratory analysis of spatial and temporal data: a systematic approach*. Springer.

Camossi, E., M. Bertolotto, and E. Bertino (2006). "A multigranular object-oriented framework supporting spatio-temporal granularity conversions". In: *International Journal of Geographical Information Science* 20.5, pp. 511–534. ISSN: 1365-8816.

Guo, D., J. Chen, A. M. MacEachren, and K. Liao (2006). "A visualization system for space-time and multivariate patterns (vis-stamp)". In: *IEEE transactions on visualization and computer graphics* 12.6, pp. 1461–1474.

Parent, C., S. Spaccapietra, and E. Zimányi (2006). "The MurMur project: Modeling and querying multi-representation spatio-temporal databases". In: *Information Systems* 31.8, pp. 733–769. ISSN: 03064379.

Schneider, M. and T. Behr (2006). "Topological relationships between complex spatial objects". In: *ACM Transactions on Database Systems (TODS)* 31.1, pp. 39–81.

Thomas, J. J. and K. A. Cook (2006). "A visual analytics agenda". In: *IEEE computer graphics and applications* 26.1, pp. 10–13.

Bédard, Y., S. Rivest, and M.-J. Proulx (2007). "Spatial. Online Analytical. Processing (SOLAP): Concepts, Architectures, and Solutions". In: *Data Warehouses and OLAP: Concepts, Architectures, and Solutions, Idea Group Inc*, pp. 298–319.

Dykes, J. and C. Brunsdon (2007). "Geographically weighted visualization: interactive graphics for scale-varying exploratory analysis". In: *IEEE Transactions on Visualization and Computer Graphics* 13.6, pp. 1161–1168.

George, B., S. Kim, and S. Shekhar (2007). "Spatio-temporal network databases and routing algorithms: A summary of results". In: *International Symposium on Spatial and Temporal Databases*. Springer, pp. 460–477.

Yuan, M. and K. S. Hornsby (2007). *Computation and visualization for understanding dynamics in geographic domains: a research agenda*. CRC Press.

Aigner, W., S. Miksch, W. Müller, H. Schumann, and C. Tominski (2008). "Visual methods for analyzing time-oriented data." In: *IEEE transactions on visualization and computer graphics* 14.1, pp. 47–60. ISSN: 1077-2626.

Camossi, E., M. Bertolotto, and T. Kechadi (2008). "Mining Spatio-Temporal Data at Different Levels of Detail". In: *The European Information Society*. Springer, pp. 225–240.

Keet, C. M. (2008). "A formal theory of granularity". PhD thesis. Citeseer.

Keim, D., G. Andrienko, J.-D. Fekete, C. Görg, J. Kohlhammer, and G. Melançon (2008). "Visual Analytics: Definition, Process, and Challenges". In: *Information Visualization*. Ed. by A. Kerren, J. Stasko, J.-D. Fekete, and C. North. Vol. 4950. Lecture Notes in Computer Science. Springer Berlin Heidelberg, pp. 154–175. ISBN: 978-3-540-70955-8.

Power, D. J. (2008). "Understanding data-driven decision support systems". In: *Information Systems Management* 25.2, pp. 149–154.

Scherr, M. (2008). "Multiple and coordinated views in information visualization". In: *Trends in Information Visualization* 38.

Zhao, J., P. Forer, and A. S. Harvey (2008). "Multi-Scale and Multi-Form Visualisation of Human Movement Patterns in the context of Space, Time and Activity: From Timeline to Ringmap". In: *Workshop on GeoVisualization of Dynamics, Movement and Change*.

Belussi, A., C. Combi, and G. Pozzani (2009). "Formal and conceptual modeling of spatio-temporal granularities". In: *Proceedings of the 2009 International Database Engineering & Applications Symposium on - IDEAS '09*. New York, New York, USA: ACM Press, p. 275. ISBN: 9781605584027.

Galton, A. (2009). "Spatial and temporal knowledge representation". In: *Earth Science Informatics* 2.3, pp. 169–187. ISSN: 1865-0473.

Hering, A. S., C. L. Bell, and M. G. Genton (2009). "Modeling spatio-temporal wildfire ignition point patterns". In: *Environmental and Ecological Statistics* 16.2, pp. 225–250.

Kechadi, M.-T., F. Ferrucci, M. Bertolotto, and S. Di Martino (2009). *Mining spatio-temporal datasets: relevance, challenges and current research directions*. INTECH Open Access Publisher.

Leung, Y (2009). *Knowledge Discovery in Spatial Data*. Advances in Spatial Science. Springer-Verlag Berlin Heidelberg. ISBN: 9783642026645. URL: http://www.google.pt/books?id=tZN-CUogHHkC.

Mennis, J. and D. Guo (2009). "Spatial data mining and geographic knowledge discovery—An introduction". In: *Computers, Environment and Urban Systems* 33.6, pp. 403–408. ISSN: 01989715.

Miller, H. J. and J Han (2009). *Geographic Data Mining and Knowledge Discovery*. Chapman & Hall/CRC data mining and knowledge discovery series. CRC Press. ISBN: 9781420073973.

Parent, C., S. Spaccapietra, C. Vangenot, and E. Zimányi (2009). "Multiple Representation Modeling." In: *Encyclopedia of Database Systems*. Ed. by L. Liu and M. T. Özsu. Springer US, pp. 1844–1849. ISBN: 978-0-387-39940-9.

Thakur, S. and T.-M. Rhyne (2009). "Data vases: 2d and 3d plots for visualizing multiple time series". In: *International Symposium on Visual Computing*. Springer, pp. 929–938.

Andrienko, G., N. Andrienko, U. Demsar, D. Dransch, J. Dykes, S. I. Fabrikant, M. Jern, M.-J. Kraak, H. Schumann, and C. Tominski (2010). "Space, time and visual analytics". In: *International Journal of Geographical Information Science* 24.10, pp. 1577–1600. ISSN: 1365-8816.

Bogorny, V. and S. Shekhar (2010). "Spatial and spatio-temporal data mining". In: *2010 IEEE International Conference on Data Mining*. IEEE, pp. 1217–1217.

Forlines, C. and K. Wittenburg (2010). "Wakame: sense making of multi-dimensional spatial-temporal data". In: *Proceedings of the International Conference on Advanced Visual Interfaces*. ACM, pp. 33–40.

Hadlak, S., C. Tominski, H.-J. Schulz, and H. Schumann (2010). "Visualization of attributed hierarchical structures in a spatiotemporal context". In: *International Journal of Geographical Information Science* 24.10, pp. 1497–1513.

Kisilevich, S., M. Krstajic, D. Keim, N. Andrienko, and G. Andrienko (2010). "Event-based analysis of people's activities and behavior using Flickr and Panoramio geotagged photo collections". In: *Information visualisation (IV), 2010 14th international conference*. IEEE, pp. 289–296.

Maciejewski, R., S. Rudolph, R. Hafen, A. Abusalah, M. Yakout, M. Ouzzani, W. S. Cleveland, S. J. Grannis, and D. S. Ebert (2010). "A visual analytics approach to understanding spatiotemporal hotspots". In: *Visualization and Computer Graphics, IEEE Transactions on* 16.2, pp. 205–220.

Malik, A., R. Maciejewski, T. F. Collins, and D. S. Ebert (2010). "Visual analytics law enforcement toolkit". In: *Technologies for Homeland Security (HST), 2010 IEEE International Conference on*. IEEE, pp. 222–228.

Møller, J. and M. Ghorbani (2010). *Second-order analysis of structured inhomogeneous spatio-temporal point processes*. Tech. rep. Department of Mathematical Sciences, Aalborg University.

Roth, R. E., K. S. Ross, B. G. Finch, W. Luo, and A. M. MacEachren (2010). "A user-centered approach for designing and developing spatiotemporal crime analysis tools". In: *Proceedings of GIScience*. Vol. 15.

Weaver, C. (2010). "Cross-filtered views for multidimensional visual analysis". In: *IEEE Transactions on Visualization and Computer Graphics* 16.2, pp. 192–204.

Aigner, W., S. Miksch, H. Schumann, and C. Tominski (2011). *Visualization of time-oriented data*. Springer Science & Business Media.

Andrienko, G., N. Andrienko, P. Bak, D. Keim, S. Kisilevich, and S. Wrobel (2011). "A conceptual framework and taxonomy of techniques for analyzing movement". In: *Journal of Visual Languages & Computing* 22.3, pp. 213–232.

Bostock, M., V. Ogievetsky, and J. Heer (2011). "D3 data-driven documents". In: *Visualization and Computer Graphics, IEEE Transactions on* 17.12, pp. 2301–2309.

De Chiara, D., V. Del Fatto, R. Laurini, M. Sebillo, and G. Vitiello (2011). "A chorem-based approach for visually analyzing spatial data". In: *Journal of Visual Languages & Computing* 22.3, pp. 173–193. ISSN: 1045926X.

MacEachren, A. M., A. Jaiswal, A. C. Robinson, S. Pezanowski, A. Savelyev, P. Mitra, X. Zhang, and J. Blanford (2011). "Senseplace2: Geotwitter analytics support for situational awareness". In: *Visual Analytics Science and Technology (VAST), 2011 IEEE Conference on*. IEEE, pp. 181–190.

Plumejeaud, C., H. Mathian, J. Gensel, and C. Grasland (2011). "Spatio-temporal analysis of territorial changes from a multi-scale perspective". In: *International Journal of Geographical Information Science* 25.10, pp. 1597–1612.

Bargiela, A. and W. Pedrycz (2012). *Granular computing: an introduction*. Vol. 717. Springer Science & Business Media.

Chae, J., D. Thom, H. Bosch, Y. Jang, R. Maciejewski, D. S. Ebert, and T. Ertl (2012). "Spatiotemporal social media analytics for abnormal event detection and examination using seasonal-trend decomposition". In: *Visual Analytics Science and Technology (VAST), 2012 IEEE Conference on*. IEEE, pp. 143–152.

Landesberger, T. von, S. Bremm, N. Andrienko, G. Andrienko, and M. Tekusova (2012). "Visual analytics methods for categoric spatio-temporal data". In: *Visual Analytics Science and Technology (VAST), 2012 IEEE Conference on*. IEEE, pp. 183–192.

Malizia, N. and E. A. Mack (2012). "Enhancing the Jacquez k nearest neighbor test for space–time interaction". In: *Statistics in medicine* 31.21, pp. 2318–2334.

Pozzani, G. and E. Zimányi (2012). "Defining spatio-temporal granularities for raster data". In: *Data Security and Security Data*. Springer, pp. 96–107.

Silva, R., J. Moura-Pires, and M. Y. Santos (2012). "Spatial Clustering in SOLAP Systems to Enhance Map Visualization". en. In: *International Journal of Data Warehousing and Mining* 8.2, pp. 23–43. ISSN: 1548-3924.

Sips, M., P. Köthur, A. Unger, H.-C. Hege, and D. Dransch (2012). "A visual analytics approach to multiscale exploration of environmental time series". In: *Visualization and Computer Graphics, IEEE Transactions on* 18.12, pp. 2899–2907.

Thom, D., H. Bosch, S. Koch, M. Wörner, and T. Ertl (2012). "Spatiotemporal anomaly detection through visual analysis of geolocated twitter messages". In: *Pacific visualization symposium (PacificVis), 2012 IEEE*. IEEE, pp. 41–48.

Tominski, C. and H.-J. Schulz (2012). "The great wall of space-time". In:

Van Ho, Q., P. Lundblad, T. Åström, and M. Jern (2012). "A web-enabled visualization toolkit for geovisual analytics". In: *Information Visualization* 11.1, pp. 22–42.

Zhang, L., A. Stoffel, M. Behrisch, S. Mittelstadt, T. Schreck, R. Pompl, S. Weber, H. Last, and D. Keim (2012). "Visual analytics for the big data era—A comparative review of state-of-the-art commercial systems". In: *Visual Analytics Science and Technology (VAST), 2012 IEEE Conference on*. IEEE, pp. 173–182.

Andrienko, G., N. Andrienko, H. Bosch, T. Ertl, G. Fuchs, P. Jankowski, and D. Thom (2013). "Thematic patterns in georeferenced tweets through space-time visual analytics". In: *Computing in Science & Engineering* 15.3, pp. 72–82.

Committee, A. o.M. D., o. A.T. S. Committee, o. M.S.T. A. Board, o. E. Division, Sciences, and Physical (2013). *National Research Council: Frontiers in Massive Data Analysis*. The National Academies Press. ISBN: 9780309287784.

Ferreira, N., J. Poco, H. T. Vo, J. Freire, and C. T. Silva (2013). "Visual exploration of big spatio-temporal urban data: A study of new york city taxi trips". In: *Visualization and Computer Graphics, IEEE Transactions on* 19.12, pp. 2149–2158.

Gabriel, E., B. Rowlingson, and P Diggle (2013). "stpp: an R package for plotting, simulating and analyzing Spatio-Temporal Point Patterns". In: *Journal of Statistical Software* 53.2, pp. 1–29.

Guo, D. and J. Wu (2013). "Understanding spatiotemporal patterns of multiple crime types with a geovisual analytics approach". In: *Crime Modeling and Mapping Using Geospatial Technologies*. Springer, pp. 367–385.

Lins, L., J. T. Klosowski, and C. Scheidegger (2013). "Nanocubes for real-time exploration of spatiotemporal datasets". In: *Visualization and Computer Graphics, IEEE Transactions on* 19.12, pp. 2456–2465.

Wang, D., W. Ding, H. Lo, M. Morabito, P. Chen, J. Salazar, and T. Stepinski (2013). "Understanding the spatial distribution of crime based on its related variables using geospatial discriminative patterns". In: *Computers, Environment and Urban Systems* 39, pp. 93–106.

Yao, J. T., A. V. Vasilakos, and W. Pedrycz (2013). "Granular computing: perspectives and challenges". In: *IEEE Transactions on Cybernetics* 43.6, pp. 1977–1989.

Alamri, S., D. Taniar, and M. Safar (2014). "A taxonomy for moving object queries in spatial databases". In: *Future Generation Comp. Syst.* 37, pp. 232–242.

Bravo, L. and M. A. Rodríguez (2014). "A Multi-granular Database Model". In: *Foundations of Information and Knowledge Systems*. Springer, pp. 344–360.

Gabriel, E. (2014). "Estimating second-order characteristics of inhomogeneous spatio-temporal point processes". In: *Methodology and Computing in Applied Probability* 16.2, pp. 411–431.

Lahouari, K., B. Jean-Yves, D. Paule-Annick, M. Hélène, and S.-M. Cécile (2014). *Représenter les dynamiques des territoires : un état des lieux, de nouveaux enjeux*. URL: http://www.map.cnrs.fr/jyb/puca/.

Laurini, R. (2014). "A conceptual framework for geographic knowledge engineering". In: *Journal of Visual Languages & Computing* 25.1, pp. 2–19.

Wang, S. and H. Yuan (2014). "Spatial data mining: a perspective of big data". In: *International Journal of Data Warehousing and Mining (IJDWM)* 10.4, pp. 50–70.

Brahim, L., K. Okba, and L. Robert (2015). "Mathematical framework for topological relationships between ribbons and regions". In: *Journal of Visual Languages & Computing* 26, pp. 66–81.

Shekhar, S., Z. Jiang, R. Y. Ali, E. Eftelioglu, X. Tang, V. Gunturi, and X. Zhou (2015). "Spatiotemporal data mining: A computational perspective". In: *ISPRS International Journal of Geo-Information* 4.4, pp. 2306–2338.

Stewart, R., J. Piburn, A. Sorokine, A. Myers, J. Moehl, and D. White (2015). "World Spatiotemporal Analytics and Mapping Project (wstamp): Discovering, Exploring, and Mapping Spatiotemporal Patterns across the World's Largest Open Soruce Data Sets". In: *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences* 2.4, p. 95.

Watson, M. C. (2015). "Time maps: A tool for visualizing many discrete events across multiple timescales". In: *Big Data (Big Data), 2015 IEEE International Conference on*. IEEE, pp. 793–800.

Cho, I., W. Dou, D. X. Wang, E. Sauda, and W. Ribarsky (2016). "VAiRoma: A visual analytics system for making sense of places, times, and events in roman history". In: *IEEE transactions on visualization and computer graphics* 22.1, pp. 210–219.

Goodwin, S., J. Dykes, A. Slingsby, and C. Turkay (2016). "Visualizing Multiple Variables Across Scale and Geography". In: *Visualization and Computer Graphics, IEEE Transactions on* 22.1, pp. 599 –608.

Li, S., S. Dragicevic, F. A. Castro, M. Sester, S. Winter, A. Coltekin, C. Pettit, B. Jiang, J. Haworth, A. Stein, et al. (2016). "Geospatial big data handling theory and methods: A review and research challenges". In: *ISPRS Journal of Photogrammetry and Remote Sensing* 115, pp. 119–133.

Robinson, A. C., D. J. Peuquet, S. Pezanowski, F. A. Hardisty, and B. Swedberg (2016). "Design and evaluation of a geovisual analytics system for uncovering patterns in spatiotemporal event data". In: *Cartography and Geographic Information Science*, pp. 1–13.

Swedberg, B. and D. Peuquet (2016). "PerSE visual analytics for calendar related spatiotemporal periodicity detection and analysis". In: *GeoInformatica*, pp. 1–21.

Zhang, J., B. Ahlbrand, A. Malik, J. Chae, Z. Min, S. Ko, and D. S. Ebert (2016). "A visual analytics framework for microblog data analysis at multiple scales of aggregation". In: *Computer Graphics Forum*. Vol. 35. 3. Wiley Online Library, pp. 441–450.

Cardoso, D., R. Alves, J. M. Pires, F. Birra, and R. A. Silva (2017). "Gisplay - Extensible Web API for Thematic Maps with WebGL". In: *Computational Science and Its Applications–ICCSA 2017*. Springer.

ArcGIS. URL: http://www.arcgis.com/ (visited on 04/01/2017).

Qlik. URL: http://www.qlik.com/us/ (visited on 04/01/2017).

Tableau Software. URL: http://www.tableau.com/ (visited on 04/01/2017).

Appendix A

The induced Relationships: Properties

Consider the relation R in the domain D:

- Symmetric Relation: $\forall x \forall y R(x,y) \rightarrow R(y,x)$
- Transitive Relation: $\forall x \forall y \forall z((R(x,y) \wedge R(y,z)) \rightarrow R(x,z))$
- Antisymmetric Relation: $\forall x \forall y R(x,y) \wedge x \neq y \rightarrow \neg R(y,x)$
- Reflexive Relation: $\forall x : xRx$

Consider a granularity G defined over the domain D, and $g_i, g_j \in G$ such that $i \in E(g_i)$ and $j \in E(g_j)$.

- Complete Relationship: $g_i \; R^C \; g_j \leftrightarrow \forall i \forall j : R(i,j)$
- Partial Relationship: $g_i \; R^P \; g_j \leftrightarrow \exists i \forall j : R(i,j) \wedge \exists j \forall i : R(i,j)$
- Weak Relationship: $g_i \; R^W \; g_j \leftrightarrow \exists i \forall j : R(i,j) \vee \exists j \forall i : R(i,j)$
- Existential Relationship: $g_i \; R^E \; g_j \leftrightarrow \exists i \exists j : R(i,j)$

An antireflexive relation R induces an antireflexive complete relation R^C The relation R^C is antisymmetric and transitive therefore the relation R^C is also antireflexive.
A reflexive relation R induces a reflexive existential relation R^E By a reflexive relation, we get: $\forall x xRx$. By an existential relation, we get: $\exists i : R(i,i)$. Naturally, the existential relation is also reflexive.
An antireflexive relation R induces an antireflexive partial relation R^P The relation R^P is antisymmetric and transitive therefore the relation R^P is also antireflexive.
A antisymmetric relation R induces antisymmetric complete relation R^C

1. $\forall x\, \forall y\, R(x,y) \rightarrow \neg R(y,x)$	
2. $\forall x\, \neg R(x,x)$	
3. $\forall i\, \forall j\, R(i,j)$	
4. $\boxed{a,b}$	
5. $R(a,b)$	$\forall$ **Elim:** 3
6. $R(a,b) \rightarrow \neg R(b,a)$	$\forall$ **Elim:** 1
7. $\neg R(b,a)$	$\rightarrow$ **Elim:** 5, 6
8. $\forall j\, \forall i\, \neg R(j,i)$	$\forall$ **Intro:** 4–7
9. $\neg R(b,a)$	$\forall$ **Elim:** 8
10. $\exists j \exists i \neg R(j,i)$	$\exists$ **Intro:** 9

Figure A.1: An antisymmetric relation induces an antisymmetric complete relation.

Note that, we want to proof $(g_i\ R^P\ g_j) \leftrightarrow (\forall j \forall i : R(j,i))$. Through DeMorgan Laws we have:

$\neg(\forall j \forall i : R(j,i))$

$(\exists j \exists i : \neg R(j,i))$

A symmetric relation R induces symmetric a complete relation R^C

1. $\forall x\, \forall y\, (R(x,y) \rightarrow R(y,x))$	
2. $\forall i\, \forall j\, R(i,j)$	
3. $\boxed{a,b}$	
4. $R(a,b)$	$\forall$ **Elim:** 2
5. $R(a,b) \rightarrow R(b,a)$	$\forall$ **Elim:** 1
6. $R(b,a)$	$\rightarrow$ **Elim:** 4, 5
7. $\forall j\, \forall i\, R(j,i)$	$\forall$ **Intro:** 3–6

Figure A.2: A symmetric relation induces a symmetric complete relation.

A symmetric relation R induces a symmetric existential relation R^E

1. $\forall x\, \forall y\, (R(x,y) \rightarrow R(y,x))$	
2. $\exists i \exists j R(i,j)$	
3. $\boxed{a,b}\ R(a,b)$	
4. $R(a,b) \rightarrow R(b,a)$	$\forall$ **Elim:** 1
5. $R(b,a)$	$\rightarrow$ **Elim:** 3, 4
6. $\exists j \exists i R(j,i)$	$\exists$ **Intro:** 5
7. $\exists j \exists i R(j,i)$	$\exists$ **Elim:** 2, 3–6

Figure A.3: A symmetric relation induces a symmetric existential relation.

A symmetric relation R induces a symmetric partial relation R^P

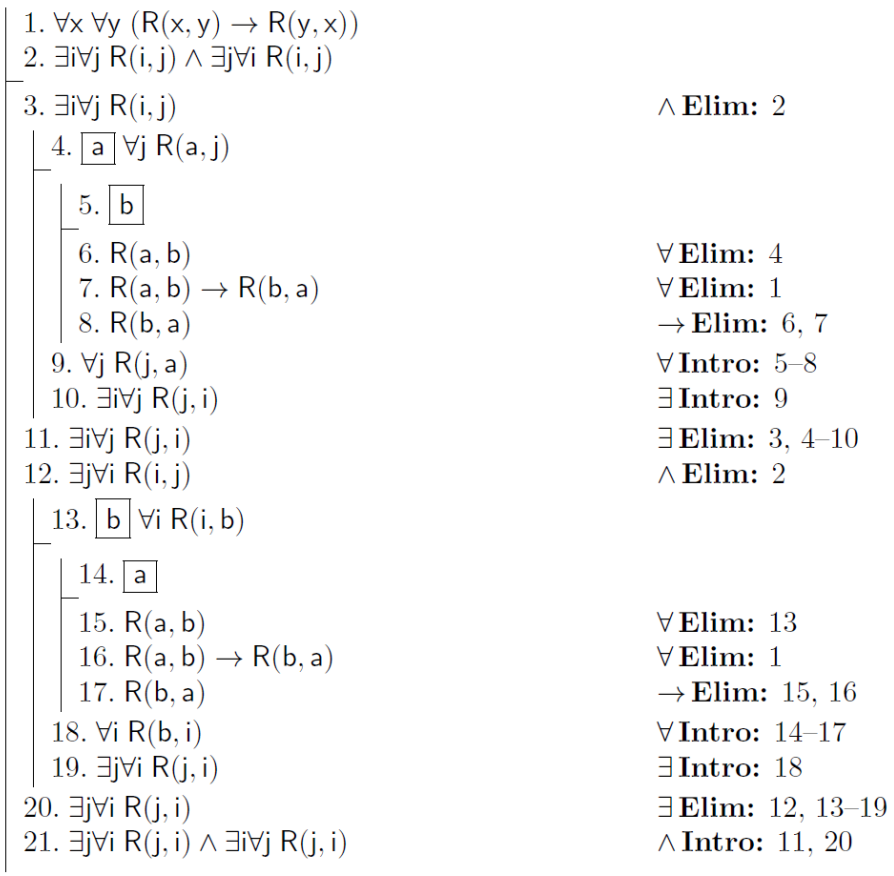


Figure A.4: A symmetric relation induces a symmetric partial relation.

A symmetric relation R induces a symmetric weak relation R^W

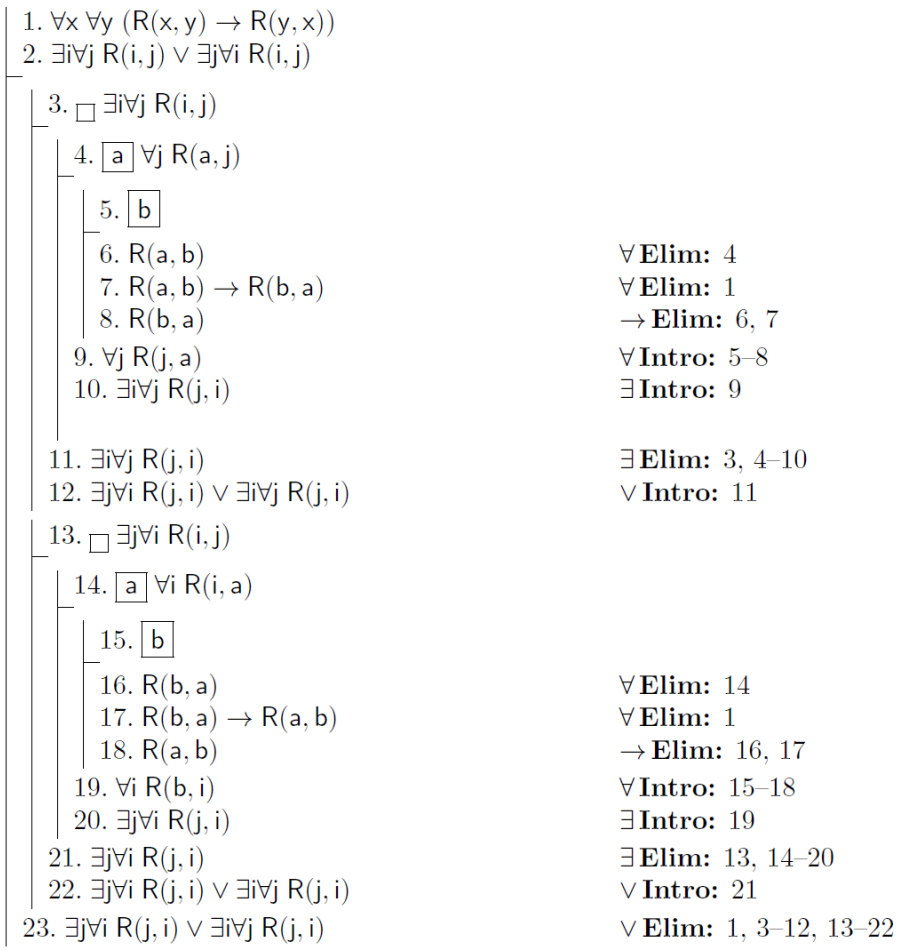


Figure A.5: A symmetric relation induces a symmetric weak relation.

An antisymmetric relation R induces an antisymmetric partial relation R^P

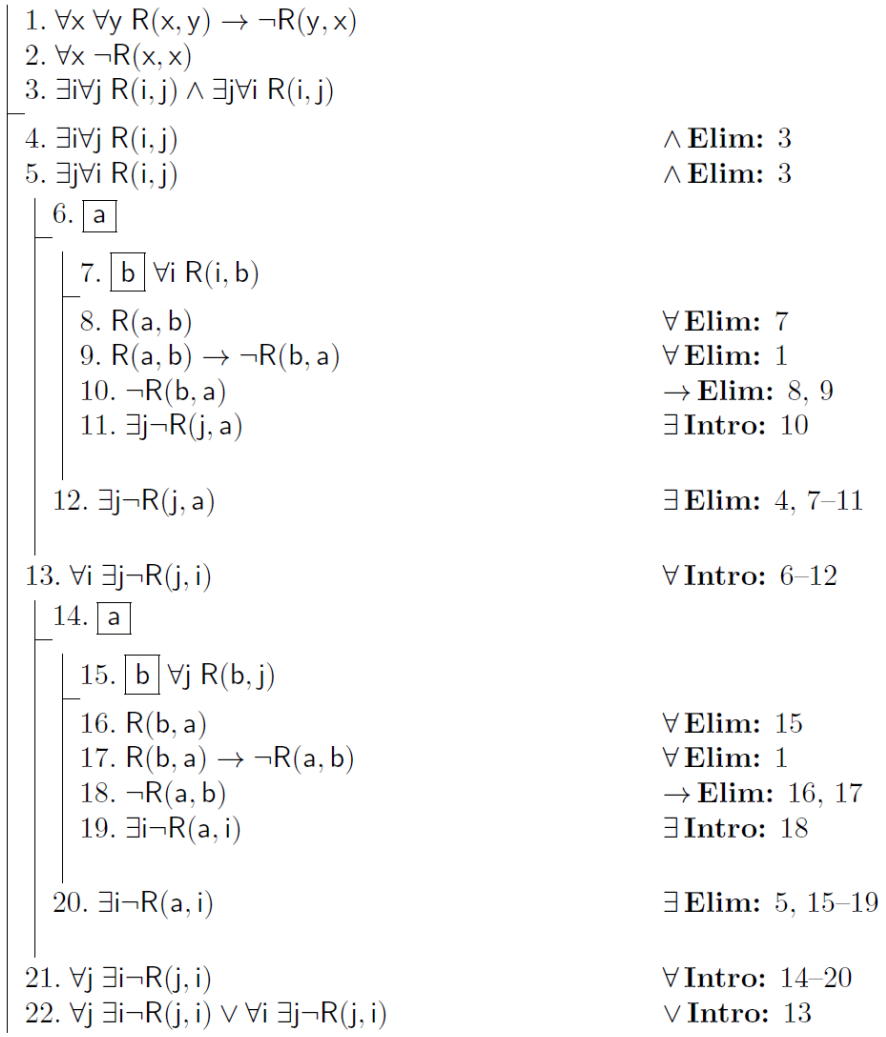


Figure A.6: An antisymmetric relation induces an antisymmetric partial relation.

Note that, we want to proof $\neg(g_j R^P g_i) \leftrightarrow \neg(\exists j \forall i : R(j,i) \wedge \exists i \forall j : R(j,i))$. Through DeMorgans Laws we have:

$$\neg(\exists j \forall i : R(j,i) \wedge \exists i \forall j : R(j,i))$$

$$(\neg\exists j \forall i : R(j,i)) \vee (\neg\exists i \forall j : R(j,i))$$

$$(\forall j \exists i : \neg R(j,i)) \vee (\forall i \exists j : \neg R(j,i))$$

Appendix B

Topological Relations on Temporal Granular Terms

In this appendix, we provide a detailed study about what happens to topological relations between temporal granular terms when these are generalized for coarser LoDs.

Throughout the following discussion, the generalization of temporal granular terms occurs between granularities related by the finer than relationship. The generalization of temporal granular terms may affect the temporal topological relationships held between pairs of atoms. On one hand, the type of relationship may change. For instance, we might have a relation between two time intervals that may turn into a relation between a time interval and a time instant. On the other hand, there are scenarios where the type of topological is kept but the actual relation (e.g., before) is changed (e.g., to equal). An overview of the possible transitions between types of topological relations is given in Figure B.1.

We start by discuss the transition expressed by the scenario one. Consider the granularities G and H such that $G \preccurlyeq H$ as defined in in Figure B.2.

Let's consider $\alpha = Interval(\alpha^-, \alpha^+)$ and $\beta = Interval(\beta^-, \beta^+)$ be intervals of time defined over a granularity G; and, let $\alpha' = Interval(\alpha'^-, \alpha'^+)$ and $\beta' = Interval(\beta'^-, \beta'^+)$ be intervals of time, generalized from α and β respectively, of a granularity H. The relation between α' and β' may be the same relation verified by the intervals of time α and β, or may be changed due to the generalization.

In the first place, when α is equal to β in any generalization scenario α' will be equal to β'. By definition, $(\alpha^-{=}\beta^-) \wedge (\alpha^+{=}\beta^+)$. Once the granularity G is finer than H then α^- and β^- will be contained by the same granule of H, and the same applies to the granules α^+ and β^+. Thus, in any generalization scenario α' will be equals to β'. The same reasoning can be applied to the meet relation. Imagine that, α meets β. We know a prior that $\alpha^+{=}\beta^-$. Therefore, α^+ and β^- will be contained by the same granule of H. Consequently, in any

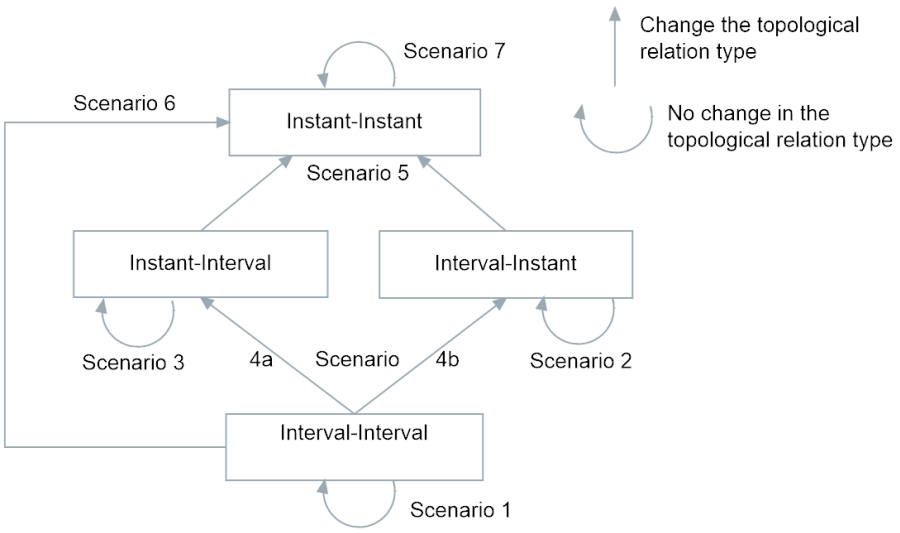


Figure B.1: Example of two granularities related by the finer-than relationship.

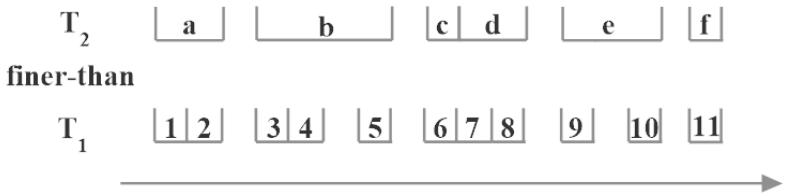


Figure B.2: Possible transitions in the relationships between pairs of temporal terms.

generalization scenario, α' will meet β'.

Furthermore, a general rule can be stated regarding the following relations: before, overlaps, starts, during, and finishes. If the endpoints of α and β that are different before generalization remain different in α' and β', i.e., after the generalization, then the relation remain unchanged. Note that, G is finer than H. Thus, when two different instants of time of G are generalized for two different instants of H then the lesser complete relationship $<^c$ [3] between them are kept. Consequently, the relation between α and β will remain between α' and β'.

Nevertheless, there are some scenarios in which the relation between two intervals of time is changed due to the generalization of them. This issue is discussed below.

Suppose that, α occurs before β. If there is $h \in H$ such that $E(\alpha^+) \subseteq E(h) \wedge E(\beta^-) \subseteq E(h)$ then α' will meet β'. For example, $\alpha = Interval(1,3)$ and $\beta = Interval\ (5,\ 6)$ such that α occurs before β. After the generalization we get: $\alpha' = Interval(a,b)$ meets $\beta' = Interval(b,c)$.

In case of α overlaps β then the relation between $\alpha^{'}$ and $\beta^{'}$ can be changed to any relation apart from the before and during relation. Consider that there are two granules $x,y \in H$ such that $E(\alpha^-) \subseteq E(x) \wedge E(\beta^-) \subseteq E(x) \wedge E(\alpha^+) \subseteq E(y) \wedge E(\beta^+) \subseteq E(y)$. In this scenario, $\alpha^{'}$ will be equals to $\beta^{'}$. This can be illustrated by considering $\alpha = Interval(3,7)$

and $\beta = Interval\ (4,\ 8)$. After the generalization we get: $\alpha' = Interval(b,d)$ equals $\beta' = Interval(b,d)$. However, if there is one granule $h \in H$ such that $E(\alpha^+) \subseteq E(h) \wedge E(\beta^-) \subseteq E(h)$ then α' will meet β'. For instance, if $\alpha = Interval(1,5)$ and $\beta = Interval\ (3,\ 6)$ then $\alpha' = Interval(a,b)$ meets $\beta' = Interval(b,c)$. Now, if there are three granules $h \in H$ such that $E(\alpha^-) \subseteq E(h) \wedge E(\beta^-) \subseteq E(h)$ then α' will start β'. For example, $\alpha = Interval(3,7)$ overlaps $\beta = Interval\ (5,\ 10)$ becomes $\alpha' = Interval(b,d)$ starts $\beta' = Interval\ (b,\ e)$. But if that granule exist such that $E(\alpha^+) \subseteq E(h) \wedge E(\beta^+) \subseteq E(h)$ then α' will be finished by β'. For example, $\alpha = Interval(6,10)$ overlaps $\beta = Interval\ (7,\ 11)$ becomes $\alpha' = Interval(c,e)$ is finished by $\beta' = Interval\ (d,\ e)$.

Let's consider the scenario in which α starts β. If there is $h \in H$ such that $E(\alpha^+) \subseteq E(h) \wedge E(\beta^+) \subseteq E(h)$ then α' will become equal to β'.

Consider that α occurs during β. For this case, if there are two granules $x, y \in H$ such that $E(\alpha^-) \subseteq E(x) \wedge E(\beta^-) \subseteq E(x) \wedge E(\alpha^+) \subseteq E(y) \wedge E(\beta^+) \subseteq E(y)$ then α' will be equals to β'. Consider the following intervals: $\alpha = Interval(4,7)$ occurs during $\beta = Interval\ (3,\ 8)$. After the generalization $\alpha' = Interval(b,d)$ is equal to $\beta' = Interval\ (b,\ d)$.

Nevertheless, if there are three granules $h \in H$ such that $E(\alpha^-) \subseteq E(h) \wedge E(\beta^-) \subseteq E(h)$ then α' will start β'. For example, $\alpha = Interval(4,6)$ during $\beta = Interval\ (3,\ 8)$ becomes $\alpha' = Interval(b,c)$ starts $\beta' = Interval\ (b,\ d)$. Contrary, if that granule exist such that $E(\alpha^+) \subseteq E(h) \wedge E(\beta^+) \subseteq E(h)$ then α' will finish β'. For example, $\alpha = Interval(6,9)$ during $\beta = Interval\ (5,\ 10)$ becomes $\alpha' = Interval(c,e)$ finishes $\beta' = Interval\ (b,\ e)$.

Finally, let's assume α finishes β. In this case, if there is $h \in H$ such that $E(\alpha^-) \subseteq E(h) \wedge E(\beta^-) \subseteq E(h)$ then α' will become equal to β'. For example, $\alpha = Interval(2,5)$ finishes $\beta = Interval\ (1,\ 5)$ becomes $\alpha' = Interval(a,b)$ equal to $\beta' = Interval\ (a,\ b)$. An overview of the previous discussion is given in Table B.1.

Table B.1: Possible transitions in the scenario 1.

	Before	Equals	Overlaps	Meets	Starts	During	Finishes
Before	√	×	×	√	×	×	×
Equals	×	√	×	×	×	×	×
Overlaps	×	√	√	√	√	×	√*
Meets	×	×	×	√	×	×	×
Starts	×	√	×	×	√	×	×
During	×	√	×	×	√	√	√
Finishes	×	√	×	×	×	×	√

So far it was assumed that the generalization of any interval of time of G results into an interval of time of H. Notice that, this study extends the results obtained by (Euzenat and Montanari 2005). Euzenat and Montanari 2005 just provides a conversion table when both temporal granules remain as intervals after their generalization for a coarser temporal granularity. However, the generalization of an interval of time of G may result

Table B.2: Possible transitions in the **scenario 4a**.

		Instant - Interval Relation				
		Before	Starts	During	Finishes	After
Interval - Interval Relation	Before	√	√	×	×	×
	Equals	Not Applicable				
	Overlaps	×	√	×	×	×
	Meets	×	√	×	×	×
	Starts	×	√	×	×	×
	During	×	√	√	√	×
	Finishes	×	×	×	√	×

into an instant of time of H which changes a relation between two intervals of time to a relation between an instant and an interval of time or the other way around (scenario two). For these cases, Euzenat's conversion table is no longer applicable.

Let's consider again $\alpha = Interval(\alpha^-, \alpha^+)$ and $\beta = Interval(\beta^-, \beta^+)$ be intervals of time defined over a granularity G. There are two ways of a relation between α and β to become a relation between an instant and an interval of time. The first one consists in α turn out to be an instant α' of H and β remains an interval of time $\beta' = Interval(\beta'^-, \beta'^+)$ of H.

In these contexts, whenever there is an granule $h \in H$ such that $E(\beta^-) \subseteq E(h) \wedge E(\alpha) \subseteq E(h)$ then any relation (except finish relation) between α and β will become α' starts β'. For example, $\alpha = Interval(3,4)$ occurs before $\beta = Interval\ (5,\ 8)$ becomes $\alpha' = b$ starts $\beta' = Interval\ (b,\ d)$. Another example can be: $\alpha = Interval(3,5)$ overlaps $\beta = Interval\ (4,\ 7)$ becomes $\alpha' = b$ starts $\beta' = Interval\ (b,\ d)$.

Suppose that, α occurs before β. If there is $h \in H$ such that $E(\beta^-) \subseteq E(h) \wedge \alpha^{'} \neq h$ then α' will occur before β'. For example, $\alpha = Interval(1,2)$ occurs before $\beta = Interval\ (3,\ 6)$. After the generalization we get: $\alpha' = a$ occurs before $\beta' = Interval(b,c)$.

Consider that, α occurs during β. If there is $x, y \in H$ such that $E(\beta^-) \subseteq E(x) \wedge E(\beta^+) \subseteq E(y) \wedge x \neq \alpha^{'} \neq y$ then α' will occur during β'. For example, $\alpha = Interval(2,5)$ occurs during $\beta = Interval\ (2,\ 6)$ becomes $\alpha' = b$ occurs during $\beta' = Interval\ (a,\ c)$. Contrary, if there is $h \in H$ such that $E(\beta^+) \subseteq E(h) \wedge \alpha^{'} = h$ then α' will finish β'. For example, $\alpha = Interval(9,10)$ occurs during $\beta = Interval\ (7,\ 11)$ becomes $\alpha' = e$ finishes $\beta' = Interval\ (d,\ e)$.

Let's assume α finishes β. In this case, if there is $h \in H$ such that $E(\beta^+) \subseteq E(h) \wedge \alpha^{'} = h$ then α' and β' will keep the relation. For example, $\alpha = Interval(4,5)$ finishes $\beta = Interval\ (1,\ 5)$ then $\alpha' = b$ also finishes $\beta' = Interval\ (a,\ b)$. Finally, if two intervals of time are equal then this discussion is not applicable because there is no scenario in which just one of them becomes an instant. Either α and β remain equal as intervals of time or as instants of time. An overview of the previous discussion is given in Table B.2.

The other possible scenario consists in α remains an interval of time $\alpha' = Interval(\alpha'^-, \alpha'^+)$ of H and β turns out to be an instant of time β' of H. In this case

Table B.3: Possible transitions in the **scenario 4b**.

		Interval.- Instant Relation				
		Before	Starts	During	Finishes	After
Interval - Interval Relation	Before	√	×	×	√*	×
	Equals	Not Applicable				
	Overlaps	×	×	×	√*	×
	Meets	×	×	×	√*	×
	Starts	Not Applicable				
	During	Not Applicable				
	Finishes	Not Applicable				

and whenever there is an granule $h \in H$ such that $E(\alpha^+) \subseteq E(h) \wedge \beta' = h$ then a before, overlaps or meets relation between α and β will become α' finished by β'. For example, $\alpha = Interval(1,3)$ occurs before $\beta = Interval\ (4,\ 5)$ becomes $\alpha' = Interval\ (a,\ b)$ finished by $\beta' = b$.

Suppose that, α occurs before β. If there is $h \in H$ such that $E(\alpha^+) \subseteq E(h) \wedge \beta' \neq h$ then α' will occur before β'. For example, $\alpha = Interval(1,3)$ occurs before $\beta = Interval\ (7,\ 8)$. After generalization, $\alpha' = Interval\ (a,\ b)$ occurs before $\beta' = d$. Regarding the relation equals, starts, during and finishes this discussion is not applicable. In these cases and by the relation definition, the extent of β contains the extent of α. In order to β turns out to be an instant β' implies that α becomes also an instant α'. An overview of the previous discussion is given in Table B.3.

Furthermore, the generalization can turn a relation between intervals of time into a relation between instants of time (**scenario 6**). Let's assume α' and β' are two instants of time of H that result from the generalization of α and β, respectively. In these circumstances, we can conclude that α' and β' will be equal in any generalization scenario except if α and β are related through the before relation. Note that, these circumstances the extent of β intersects the extent of α. As a result, in order to α and β turn out to be instants implies that α' and β' are equal.

When α occurs before β, after the generalization, α' and β' can also be equal or the before relation is "maintained". If there is $h \in H$ such that $E(\alpha^-) \subseteq E(h) \wedge E(\beta^+) \subseteq E(h)$ then α' will occur before β'.

Until now, the discussion about the generalization of temporal terms and temporal relations has its starting point from the generalization of two intervals of time. Now, let's consider α and $\beta = Interval(\beta^-, \beta^+)$ be an instant and an interval of time defined over a granularity G, correspondingly; and, let α' and $\beta' = Interval(\beta'^-, \beta'^+)$ be an instant and an interval of time, generalized from α and β respectively, of a granularity H (**scenario 4**).

A general rule can be stated regarding the relations between an instant and an interval of time: if the granules involved (α and the endpoints of β) that are different before generalization remain different in α' and β', i.e., after the generalization, then the relation remain unchanged. The rationale is the same as it was in the generalization between

Table B.4: Possible transitions in the **scenario 3**.

		Instant - Interval Relation				
		Before	Starts	During	Finishes	After
Instant- Interval Relation	Before	$\surd$	$\surd$	×	×	×
	Starts	×	$\surd$	×	×	×
	During	×	$\surd$	$\surd$	$\surd$	×
	Finishes	×	×	×	$\surd$	×
	After	×	×	×	$\surd^*$	$\surd$

Table B.5: Possible transitions in the **scenario 2**.

		Interval - Instant Relation				
		Before	Starts^{-1}	During^{-1}	Finishes^{-1}	After
Interval- Instant Relation	Before	$\surd$	×	×	$\surd^*$	×
	Starts^{-1}	×	$\surd$	×	×	×
	During^{-1}	×	$\surd$	$\surd$	$\surd$	×
	Finishes^{-1}	×	×	×	$\surd$	×
	After	×	×	×	$\surd^*$	×

intervals of time. This is also applicable in case of interval-instant relations.

There are a few scenarios in which the relation between α and β is different from the relation between α' and β'. Suppose that, α occurs before β. If there is $h \in H$ such that $E(\alpha) \subseteq E(h) \ \wedge E(\beta^-) \subseteq E(h)$ then α' will occur before β'. Now, consider that α occurs during β. If there is $h \in H$ such that $E(\alpha) \subseteq E(h) \ \wedge E(\beta^-) \subseteq E(h)$ then α' will starts β'. Contrary, if there is $h \in H$ such that $E(\alpha) \subseteq E(h) \ \wedge E(\beta^+) \subseteq E(h)$ then α' will finish β'. On the other hand, let's consider α occurs after β. If there is $h \in H$ such that $E(\alpha) \subseteq E(h) \ \wedge E(\beta^+) \subseteq E(h)$ then α' will be finished by β'.An overview of the previous discussion is given in Table B.4.

A similar discussion can be made if we consider α as an interval of time and β an instant of time (**scenario 3**). An overview of the possible transitions is displayed in Table B.5.

In same the way, the generalization can turn a relation between intervals of time into a relation between instants of time also a relation between an instant and an interval of time (or vice-versa) can become a relation between two instants of time (**scenario 5**). In case of α and β are related through the start, during or finishes relation, in any generalization scenario α' and β' will be equal. The reason is similar to the one exposed in the case of intervals of time. If α occurs before or after β then α' may keep occur before, or after respectively β', or be equal. The circumstances in which these changes occurs are similar to the scenario four.

Last but not least, when two different instants of time of G, α and β, are generalized for two instants of H, α' and β' (**scenario 7**), the relationship between α and β are kept if and only if $E\left(\alpha^{'}\right) \cap E\left(\beta^{'}\right) = \emptyset$. Otherwise both instants of time become the same (at the granularity H). Furthermore, if two instants of time of G are equal, after the generalization,

they remain equal.

Appendix C

Abstracts Implemented

Measure	Description	Global	Spatial	Temporal	Properties
Collision Rate (%)	Percentage of spatiotemporal granules with events, where atom collisions exists	•	•	•	
Occupation Rate (%)	Percentage of granules with events	•	•	•	
Granular Mantel Bounded and Normalized	Measures the spatiotemporal interaction	•			⊡
Frequency Rate (%)	Percentage of events happened in granules		•	•	

Legend:

⊡ Neighbourhood dependent ⊛ Semantic dependent

Measure	Description	Global	Spatial	Temporal	Properties
Bray-Curtis Similarity for Atoms	Calculates the similarity based on the counts of atoms, between consecutive temporal grains		•		⊡
Bray-Curtis Similarity for Synthesis	Calculates the similarity based on the number of granular synthesis, between consecutive temporal grains		•		⊡
Correlation Index For Atoms	Correlation between the number of atoms of consecutive temporal grains		•		⊡
Correlation Index For Synthesis	Correlation between the number of granular synthesis of consecutive temporal grains		•		⊡
Dice Similarity (Binary)	Dice index (event / no event) between consecutive temporal grains		•		⊡
Jaccard Similarity (Binary)	Jaccard index (event / no event) between consecutive temporal grains		•		⊡

Measure	Description	Global	Spatial	Temporal	Properties
Gower Similarity (Binary)	Similarity (event / no event) between consecutive temporal grains		•		⊡
Moran's I	Calculates the spatial autocorrelation among nearby locations, given a domain specific variable		•		⊡ ⊛
Nearest Neighbor (NN)	Measures the level of clustering		•	•	⊡
Z-score Nearest Neighbor (z-NN)	Measures the z-score of the level of NN		•	•	⊡
Spatial Scope	Measures the spatial extent		•		⊡
Spatial Consecutive Distance between Centers of Mass	Measures the distance between consecutive centers of mass		•		⊡
Center's Mass Positioning	Measures the position of the centers of mass		•	•	⊡
Reduction Rate (%)	Measures the reduction of atoms used	•			
Average Atoms in spatiotemporal granules (%)	Measures the average of atoms indexed by spatiotemporal granules	•			

APPENDIX D

Granular Mantel Bounded and Normalized

Popular methods that measure spatiotemporal interaction are Knox and Bartlett 1964, Mantel 1967, Jacquez 1996 k Nearest Neighbor. The purpose of these tests is to have a measure about the presence or absence of spatiotemporal clustering pattern or other pattern that involves spatiotemporal interaction like the contagious process.

The previous methods have been used on spatiotemporal events in order to determine to if the events are "interacting". In the end, the mentioned tests check whether events are close to each other in space and in time. However, there are differences among their computation.

The Knox statistic is calculated as the total number of event pairs where the spatial and temporal distances (d^s_{ij} and d^t_{ij}, respectively) between pairs are within the specified thresholds (α and β) (see Equation D.1) If interaction is present, the test statistic will be large. Notice that, n is the number of events.

$$
\begin{aligned}
Knox &= \sum_{i=1}^{n}\sum_{j=1}^{n} a^s_{ij} a^t_{ij} \\
a^s_{ij} &= \left\{ \begin{matrix} 1, d^s_{ij} < \alpha \\ 0, otherwise \end{matrix} \right\} \\
a^t_{ij} &= \left\{ \begin{matrix} 1, d^t_{ij} < \beta \\ 0, otherwise \end{matrix} \right\}
\end{aligned}
\tag{D.1}
$$

The Mantel test keeps the distance information discarded by the Knox test. There are two versions of the Mantel test statistic: (*i*) an unstandardized; (*ii*) a standardized version. The unstandardized Mantel statistic is calculated by summing the product of the spatial d^s_{ij} and temporal distances d^t_{ij} between all event pairs (see Equation D.2). Notice that, Mantel introduces a constant c to the distance to prevent multiplication by zero.

$$Mantel_{unnormalized} = \sum_{i=1}^{n}\sum_{j=1}^{n} = (d_{ij}^{s} + c)(d_{ij}^{t} + c) \tag{D.2}$$

The standardized test statistic is calculated by measuring the correlation the spatial and temporal distance matrices:

$$Mantel_{normalized} = \frac{1}{n^2 - n - 1}\sum_{i=1}^{n}\sum_{j=1}^{n}\left[\frac{d_{ij}^{s} - \bar{d}^{s}}{\sigma_{d^s}}\right]\left[\frac{d_{ij}^{t} - \bar{d}^{t}}{\sigma_{d^t}}\right] \tag{D.3}$$

where $\bar{d}^{s}$ and $\bar{d}^{s}$ are the average distance in space and time, correspondigly; and σ_{d^s} and σ_{d^t} refers to standard deviations, for distances in space and time, respectively. This test statistic is the Pearson product-moment correlation coefficient. Therefore, the value falls in the range of -1 to +1, where being close to -1 indicates strong negative correlation, +1 means strong positive correlation, and 0 indicates no correlation.

Instead of using the distance in space and time to determine proximity (like the Knox test) the Jacquez test employs a nearest neighbor distance approach. The statistic is calculated as the number of event pairs that are within the set of k nearest neighbors from each other in both space and time, which is expressed formally as:

$$\begin{aligned} Jacquez &= \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ijk}^{s} a_{ijk}^{t} \\ a_{ijk}^{s} &= \left\{\begin{matrix} 1, \text{if event j is a k nearest neighbor of event i in space} \\ 0, \text{otherwise} \end{matrix}\right\} \\ a_{ijk}^{t} &= \left\{\begin{matrix} 1, \text{if event j is a k nearest neighbor of event i in time} \\ 0, \text{otherwise} \end{matrix}\right\} \end{aligned} \tag{D.4}$$

We aim to measure spatiotemporal interaction among spatiotemporal events described at some spatiotemporal LoD γ instead of raw spatiotemporal events. This arise two requirements. One hand the measure must be comparable among LoDs of the *event* predicate, and on the other hand, the measure must handle granular syntheses. Obviously, none of the tests detailed were designed to handle granular syntheses. However, let's discuss whether they might provide comparable values among LoDs or not.

The Mantel test (both versions) work with all events, and therefore, this test cannot discover changes in the pattern of correlation at different distances (i.e., LoDs). To discuss Knox and Jacquez test, let' us mention a particularity when we have spatiotemporal events modeled through the granularities-based model.

In some LoD of the *event* predicate, we have granular syntheses "interacting" with each other. As long as we move to coarser LoDs, the co-occurrence of granular syntheses in spatiotemporal grains increases. As a result, as long as we move to coarser LoDs, the probability of pair of granular syntheses being at zero distance between each other also increases. Consequently, the Knox and Jacquez values tend to increase as we consider

coarser LoDs. This property make Knox and Jacquez not comparable among LoDs from our perspective.

To meet this need, we introduce a new approach to measure the spatiotemporal interaction among granular syntheses in some LoD of the *event* predicate. This is expressed formally as:

$$GMBN = \frac{1}{sum(f_j)} \sum_{i=1}^{n} \sum_{j=1}^{n} (a_{ij}^s + c_s)(a_{ij}^t + c_t) * f_j$$
$$a_{ij}^s = \begin{Bmatrix} \frac{d_{ij}^s}{\alpha}, d_{ij}^s < \alpha \\ 0, otherwise \end{Bmatrix} \tag{D.5}$$
$$a_{ij}^t = \begin{Bmatrix} \frac{d_{ij}^t}{\beta}, d_{ij}^t < \beta \\ 0, otherwise \end{Bmatrix}$$

Our approach is an extension of the Mantel in order to take into account the **granular context**. This way, the proposed test is calculated by summing the product of the spatial a_{ij}^s, temporal distances a_{ij}^t and the number of events at the granular synthesis j, i.e. f_j between all granular synthesis pairs (see Equation D.5). Notice that, similarly to Mantel, there is a constant c_t and c_s to prevent multiplication by zero.

Furthermore, in our case, the test is **bounded** as we are just considering the neighbors within a spatial distance of α and a temporal distance of β. Moreover, in our case, the test is also **normalized** as the distances a_{ij}^s and a_{ij}^t are being normalized by α and β, correspondingly. This way, both space distances and temporal distances are placed between 0 and 1, and contribute equally to the end result. Finally, the end result is normalized by the summing of all f_j. For these reasons, we called the proposed test **the Granular Mantel Bounded and Normalized (GMBN)**.